From Text to Landscape: Extraction of Landscape Concepts through the Resolution of Ambiguity and Vagueness present in Descriptions of Natural Landscapes.

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Curdin Derungs

von

Brigels GR

Promotionskommitee

Prof. Dr. Ross Purves (Leitung der Dissertation)

Prof. Dr. Robert Weibel

Dr. Bettina Waldvogel

Martin Hägeli

Zürich 2014

Zusammenfassung

Wie beschreiben Menschen ihre unmittelbare Umgebung? Diese Frage ist zentral für viele Aufgaben von sozialer Relevanz. Beispiele sind die *Raumplanung*, das *Ressourcenmanagement* oder *Krisenintervention*. Für diese Aufgaben gilt, dass räumliche Information wichtig ist, die widerspiegelt wie der Menschen den Raum versteht. Falls die Information nicht dem menschlichen Verständnis des Raumes entspricht, ist sie nicht nützlich und kann zu falschen Entscheidungen führen.

Landschaftsbeschreibungen enthalten viele Unsicherheiten und sind darum eine Herausforderung für die Geographie. Die meisten Unsicherheiten gründen auf der menschlichen Wahrnehmung. Menschen haben unterschiedliche Konzepte von der gleichen Landschaft und nutzen darum unterschiedliche Worte um sie zu beschreiben. Das ist insbesondere wahr wenn Menschen aus unterschiedlichen Kultur- und Sprachgruppen verglichen werden. Eine andere Ursache von Unsicherheit hängt mit den Objekten zusammen die genutzt werden, um Landschaften zu beschreiben. Beispiele von solchen Objekten sind *Berg, Tal, Hügel, Fluss* oder *Wald*. Ein Berg ist beispielsweise weder rein natürlich, noch ist er eindeutig Mensch-gemacht. Die stoffliche Basis von geographischen Objekten ist meist natürlich, so zum Beispiel der Stein, der dem Berg seine Form gibt. Einen Ausschnitt des kontinuierlichen Verlaufes der Erdoberfläche aber als individuelle Objekte wahrzunehmen ist menschlich. Dieser Umstand gestaltet sowohl die semantische, als auch die räumliche Definition von Landschaftsobjekten als äusserst schwierig.

Die beschriebenen Unsicherheiten werden oft mit *Vagheit* bezeichnet. Für den Menschen und seinen Alltag ist Vagheit kaum hinderlich. Im Gegenteil: Vagheit ist eine wichtige Voraussetzung für menschliche Kommunikation. „Ich war am Wochenende in den Bergen!“ wird vom Gegenüber wohl problemlos verstanden. Das Verwenden eines vagen Konzeptes, hier *Berg*, garantiert, dass der Satz nicht zu kompliziert oder umständlich wird. Vagheit ist aber dann eine Herausforderung, wenn wir Landschaftsbeschreibungen im Computer speichern möchten. Klassische *Geographische Informations Systeme* sind für präzise Information geschaffen. Grenzen haben beispielsweise oft abrupten Charakter, definiert durch scharfe Linien, und Attributwerte sind oft numerisch oder kategorisch. Zudem ist es nicht üblich, das gleiche Objekt mehrmals zu speichern, um dadurch unterschiedliche menschliche Wahrnehmungen abzudecken.

Menschliche Landschaftskonzepte und deren Vagheit zu erfassen ist bereits Gegenstand geographischer Forschung. In der Ethnophysiography werden beispielsweise Menschen nach ihrem Landschaftskonzept

befragt. Die Befragung findet oft im Feld statt und bei den befragten Personen handelt es sich meist um Angehörige indigener Völker. Solche Forschung erfasst Landschaftskonzepte mit beachtlicher räumlicher Auflösung. Der offensichtliche Nachteil von ethnophysiographischer Forschung ist der grosse Aufwand zur Informationsgewinnung und damit verbunden auch die oft nur limitierte räumliche Abdeckung.

In dieser Arbeit nutzen wir schriftliche, digitalisierte Landschaftsbeschreibungen, um damit der räumlichen Limitierung von ethnographischer Forschung entgegenzuwirken. Die Nutzung von geographischer Information aus unstrukturierten Beschreibungen bedingt aber, dass wir die Information in einem ersten Schritt zu extrahieren haben. Werkzeuge und Herangehensweisen dafür finden sich in einer Vielzahl von Disziplinen. Beispiele dafür sind *Digital Humanities*, *Literary GIS*, *Geographic Information Retrieval* (GIR) und Arbeiten mit *User Generated Contents*.

Räumliches Referenzieren von Landschaftsbeschreibungen. In einem ersten Schritt weisen wir eine Kollektion von mehreren hundert Büchern die Landschaftsbeschreibungen enthalten dem geographischen Raum zu. GIR bietet Möglichkeiten und Algorithmen und dies zu bewerkstelligen, hauptsächlich indem Ortsnamen in den Beschreibungen erkannt und mit geographischen Koordinaten assoziiert werden. Die meisten Arbeiten in GIR arbeiten mit Textdokumenten die sich auf Länder, Städte, Gemeinden oder Kantone beziehen, also relative bekannte Orte. Das Referenzieren von Landschaftsbeschreibungen ist hingegen eine grosse Herausforderung, da aufgrund der räumlichen Detailliertheit der Beschreibungen viel Ortsnamen vorkommen die nur wenig bekannt sind. Um solche spezifischen und wenig bekannten Ortsnamen, wie beispielsweise die Namen von Bergen, Hügeln oder Fluren berücksichtigen zu können mussten wir eine neue Methode entwickeln, die unabhängig von der Art der Ortsnamen funktioniert. Wir haben dazu die Annahme getroffen, dass Landschaftscharakteristiken wie die Topographie genutzt werden können, um das Erkennen von Ortsnamen zu unterstützen. Eine Evaluation dieser neuen Methode hat gezeigt, dass sich damit die Qualität der Resultate signifikant verbessert. Das heisst, wir können Landschaftsbeschreibungen genauer dem geographischen Raum zuordnen als dies mit klassischen Algorithmen möglich ist. Zudem können wir zeigen, dass eine genaue räumliche Referenzierung von Landschaftsbeschreibungen der Schlüssel für das korrekte Beantworten von Suchanfragen mit räumlicher Komponente ist.

Ein Produkt das von räumlich referenzierten Landschaftsbeschreibungen abgeleitet werden kann sind Kartierungen, welche die räumliche Verteilung von mehr als hundert Büchern zeigen. Im Literary GIS wird argumentiert, dass solche Karten eine Ergänzung zum klassischen Lesen von Texten sind. Daraus können Informationen gewonnen werden, die durch das Lesen der Texte nicht oder nur sehr aufwändig erlangt werden können. Literary GIS nutzt zwar die Kartierung von Texten als linguistisches

Analysewerkzeug, die Kartierung wird dabei aber manuell erfasst. Bei uns funktioniert die Kartierung hingegen automatisch. Damit können wir grosse Datenmengen bearbeiten und zum Beispiel zeigen, wie sich der räumliche Fokus von Landschaftsbeschreibungen über die letzten 150 Jahre geändert hat.

Landschaftsinformation. In einem zweiten Schritt bewegen wir uns von (geographisch referenzierten) Landschaftsbeschreibungen hin zur Extraktion und Speicherung von Landschaftsinformation. Als Landschaftsinformation bezeichnen wir die Art und Weise, wie Landschaftsobjekte in Beschreibungen verwendet werden. Das Erkennen von Landschaftsobjekten ist durch einen vorgelagerten Arbeitsschritt gewährleistet. Eine Gruppe von Freiwilligen hat dabei geholfen, aus einer Liste von 1500 häufigen Substativen diejenigen zu markieren, welche natürliche Landschaften beschreiben. Dabei kommt ein detailliertes Regelwerke zur Anwendung. Das Resultat dieses Arbeitsschrittes ist eine Liste mit 94 Landschaftsobjekten. Die (relative) Häufigkeit dieser Landschaftsobjekte in Beschreibungen können wir messen und als lokale Landschaftsinformation speichern. Eine solche Speicherung von Landschaftsinformation bietet die Möglichkeit, dass Vagheit in Landschaftsbeschreibungen erhalten bleibt, ohne dass dies die Datenspeicherung verunmöglichen würde. Der Vergleich von Landschaftsinformation wie sie an unterschiedlichen Orten gespeichert wird kann qualitativ und quantitativ untersucht werden.

Beiträge. Die Beiträge dieser Dissertation zum Stand der Forschung sind auf eine Reihe von Bereichen verteilt. Ein methodischer Beitrag zu GIR und Literary GIS besteht aus einer neuen Technik um Landschaftsbeschreibungen zu referenzieren. Dies war bis anhin nur mit limitierter Präzision möglich oder mit grossem Aufwand verbunden. In diesem Zusammenhang können wir beispielsweise zeigen, dass Suchmaschinen zum Prozessieren von räumlichen Suchen mit detaillierter Auflösung geographische Information berücksichtigen müssen. Das konnte zuvor noch nie so klar gezeigt werden. Durch das Strukturieren von Landschaftinformation konnten wir einen weiteren methodischen Beitrag leisten, dieses Mal zum Thema *Kompatibilität von GIS zum prozessieren von vager menschliche Information*. Die lokale Landschaftsinformation die wir aus Texten extrahiert haben ist in einer Reihe von Untersuchungen zur Anwendung gekommen: Einerseits wurde der Stand von ethnophyiographischer Forschung um eine Methode erweitert mit der menschliche Landschaftsbeschreibungen für grosse (Zeit-)Räume gewonnen werden kann. Der Detailierungsgrad der gewonnen Landschaftsinformation erlaubt sowohl qualitative wie auch den quantitative Vergleiche mit alternativen Informationsquellen. Wir können beispielsweise zeigen, dass die Variation von Landschaftsbeschreibungen in der Schweiz an lokale topographische Charakteristiken gekoppelt ist. Dies ist insbesondere interessant da es Möglichkeiten zeigt, wie lokale Landschaftsinformation aus physikalischen Parametern abgeleitet werden kann, was sich im Kontext von lokalem Informationsbedürfnis für Internetsuchen als spannend erweisen könnte.

Summary

How do local people describe landscapes? This question is crucial for tackling many tasks of social relevance such as *land use planning, natural resource management* and *crisis intervention*. For all of these it is of crucial importance to have spatial information available, and in particular information that reflects how individuals conceptualize space, in order to make the appropriate decisions.

From a geographic perspective the relevance of the question as to how people describe landscape is additionally challenging since landscape descriptions are the source of numerous uncertainties. Most of these uncertainties are the result of human perception. For example, different actors may have different concepts of the same landscape and thus describe it using different words. This is particularly true for people from different cultures or language groups. Furthermore, the descriptions of the objects making up the landscape are also prone to uncertainties. Thus, for example, objects such as mountains, valleys, rivers and forests are difficult to define semantically and spatially. For instance, a mountain is neither a product of natural selection, nor is it purely artificial. The physical basis of a mountain, such as the rock from which it is formed, is natural, whereas the delineation of its extent from the earth's surface is clearly a human, or artificial, product. Such uncertainties are often synonymously related to vagueness. We successfully deal with vagueness in everyday situations without any difficulty. Indeed, vagueness is inherent to natural language and a building block of successful communication. The statement "I spent the weekend in the mountains!" in a conversation would be unremarkable and the use of the vague concept *mountain* guarantees that the statement is not cluttered with irrelevant details. However, vagueness is a challenge if landscape descriptions are to be stored in a computer. Typical Geographic Information Systems are well suited for storing and analyzing precise information, with boundaries being sharp and attributes often having numeric values. Furthermore, it would not be standard practice to represent several versions of the same landscape object in order to capture vagueness in terms of variations in human perception.

Capturing information on how landscapes are described and the precise characterization of vagueness in such descriptions has long been a goal of geographic research. In ethnophysiography, for example, local people are asked to describe key landscape concepts. Such inquiries usually take place in the field, in the form of interviews or field walks. The interviewees are often indigenous people from ethnic groups distributed all over the globe. Ethnophysiographic research thus gathers information about landscape concepts at detailed local scales - at the obvious cost, however, of intensive efforts in the collection of the information and often limited spatial coverage.

In this thesis we aim to explore a new source of information for landscape descriptions and thereby address some of the limiting factors of ethnographic or field based approaches. We use written landscape descriptions contained in large compilations of digitized books. However, using geographic information from unstructured natural language sources requires us to firstly make the information explicit. Tools and approaches that are associated with this task are described in a number of disciplines, such as *digital humanities*, *literary GIS*, *geographic information retrieval* (GIR) and recent work with *user generated content*.

Linking Landscape Descriptions to Spatial Footprints. In a first step we aim to link some hundred volumes of text containing landscape descriptions to spatial footprints. The GIR literature offers a number of approaches for performing this task, mainly through recognizing and associating place names in text with geographic coordinates. However, landscape descriptions constitute a particular challenge to the state of the art in GIR, mainly because of the fine spatial granularity of the descriptions. Previous work in GIR has mainly focused on descriptions with place names referring to cities or communities. In order to process detailed descriptions, containing references to mountains, hills or other natural features, we introduced a new heuristic independent from the type of place name. We thus assume that particularities of place names that refer to geographic objects can be characterized using topographic information and that such information is useful for correctly recognizing and referencing place names in text. An evaluation of our heuristics shows that our final product, consisting of the spatial footprints of some 10,000 landscape descriptions, is significantly more precise compared to a state of the art baseline. Additionally, we applied our results to a spatial information retrieval task and compared it with traditional information retrieval, such as for instance performed by commercial search engines. We can thus show that for the retrieval of relevant results from detailed spatial information and for detailed queries it is crucial to use geographic intelligence. State of the art information retrieval cannot sufficiently cope with this task.

A second product from the linking of landscape descriptions to spatial footprints is a map that represents the spatial distribution and the focus of some hundred books. Literary GIS argues that such maps are an important addition to traditional close reading, since they offers insights on the content of books that cannot be reached through a close reading. Thus we can, for instance, show how the spatial footprints of landscape descriptions have changed over the last 150 years.

Landscape Information. In a second step we move from georeferenced landscape descriptions towards the extraction and storage of explicit landscape information. Landscape information is approximated from particular uses of geographic objects in descriptions. The recognition of geographic objects in text is

guaranteed through a preprocessing step, where a group of volunteers annotated some 1500 frequent nouns from descriptions for filtering out geographic objects according to a set of annotation rules. Thus, we retained a set of 94 geographic objects. The (relative) frequencies of the use of these geographic objects in descriptions are taken as a proxy for deducing local landscape information. This methodology for extracting and storing landscape information allows us to capture some of the vagueness in landscape descriptions. Landscape information gathered from different landscapes can either be qualitatively or quantitatively compared. Qualitative comparisons focus on the use of geographic objects, whereas in quantitative comparisons numeric values from the frequency distribution of geographic objects are used to apply statistics.

The work in this thesis is associated with contributions that relate to different scientific domains. The new approach for linking landscape descriptions to spatial footprints can be considered a methodological contribution to GIR and literary GIS. Previous to our approach, this task was resolved with only limited spatial precision or it was very time consuming. In the same context, we could show that for correctly processing spatial queries of fine spatial resolution, a search engine necessarily needs to incorporate geographic information. This has never been shown before. A second methodological contribution is represented by our approach for extracting and structuring geographic information from landscape descriptions. This time the contribution is embedded in the context of compatibility of GIS for vague human sourced information. We used the local landscape information in a series of applications and could thus show that we contribute to the state of the art in ethnophysiographic research, in particular by extending the spatial and temporal coverage. The retrieved landscape information is comprehensive enough to be related to alternative sources information. We could thus show that landscape descriptions are statistically related to local topographic characteristics. This could be relevant for *local search* applications in the internet. Lacking local information could be approximated through local physical measurements.

Acknowledgments

I'd like to specially thank the following institutions, groups and, most importantly, people:

This thesis was granted by the **Swiss Science Foundation**, under contract 200021-126659, and carried out at the **Geography Department of the University of Zurich**. To both I owe my debts for supporting my work through money, infrastructure and interest.

The thesis was accomplished through cooperation with the **Swiss Federal Institute for Forest, Snow and Landscape Research** (WSL). The chance of having two work places and two research groups to relate to clearly broadened my horizon and introduced a welcome change to my working week. In particular I'd like to say my thanks to my contact persons at WSL, **Martin Hägeli** and **Bettina Waldvogel**.

I would particularly like to thank **Ross Purves**, my main supervisor. In all these years of countless collaboration I have always highly respected him as a person and mentor. We had countless arguments on important and not so important things. In the end, we sometimes agreed. The collaboration with Ross, from time to time, was challenging but always enriching, efficient and most often fun!

I owe debts to all my **colleagues** at the University of Zurich, many of whom are directly responsible for my having highly enjoyed the time of my PhD. One person that requires to be separately mentioned is **Christian Gschwend**. If you ever have to share your office with someone, I suggest choosing Christian. He is a well-balanced mixture between serenity and benevolence, and yet very funny.

David Mark spent many hours explaining the bigger picture to me, where everything has its proper place, whether it be geography, psychology or linguistics. I took great advantage from these hours, although I sometimes felt a bit puzzled and often hoped not to make an all too puzzled impression! Anyways, I'd like to thank David for all his support!

I always felt the strong support from my **family**. At the same time, they gave me all the freedom to independently choose from early on in my life (my mother was once told by a chiromancer that there is no point in keeping me on a short leash… she obviously took it seriously). I'd like to thank them from the bottom of my heart, since this is the true foundation for me to make my way.

Last, but most of all, I'd like to thank **Esther** for everything that was, is and might come. What she is to me goes far beyond the supportive role she played in this thesis. Still, I would not have had the strength to stay motivated without her. But as I said, that's pretty much true for everything in my life!!

Contents

Chapter 1 Introduction

Increasing volumes of data are digitally available with, for instance, more than 20 million books added to Google Books over the last decade. This is digital information in the form of unstructured text, which could be considered in scientific domains that traditionally focused on analogous data, such as interviews, empirical experiments or field walks. In social sciences this is reflected by the rise of the topic *digital humanities* (Berry 2012) (Figure 1).

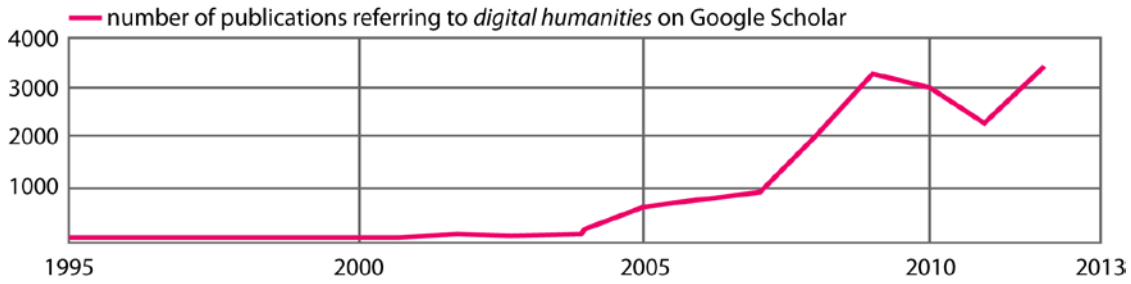


Figure 1. Rise of the topic *digital humanities* in scientific publications.

Moretti (2007) points out that in traditional humanities a collection of 200 novels on 19th century British literature was considered extensive, but is still less than 1% of the novels published in this period. Close reading of all twenty or thirty thousand British novels published in the 19th century is not feasible, as it would require more than a century for one person to do. Margret Cohen (1999) calls this gap between collections analyzed and documents theoretically available the *great unread*. Digital humanities can be seen as a reaction to the great unread by answering research questions from the humanities through the automatic processing of large digital data, often in the form of digitized books.

However, books are written in natural language and thus have unstructured content. Structure, in terms of explicit information, has first to be imposed in order to deduce interpretations. An impressive example of imposing structure onto digitized books is the *Google Books Ngram Viewer*, where the evolution of an arbitrary word or topic can be followed over time. The information is retrieved from a corpus consisting of over 20 million linguistically parsed text documents, mostly books, published between 1500 and 2008 (Michel *et al.* 2011). Figure 2 contains two examples of plots, as generated by the Google Books Ngram Viewer, with the term *mountain* being consistently used over time. Other terms, such as *computer*, clearly reflect societal trends.

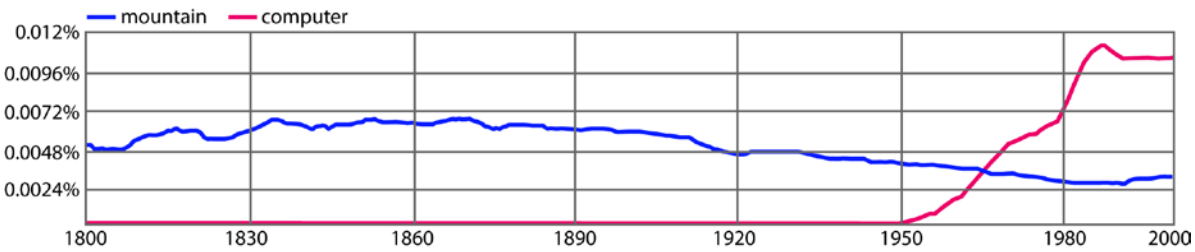


Figure 2. Temporal plots for the terms *mountain* and *computer* retrieved using the Google Ngram Viewer.

The role of geography, in the context of the availability of large digital libraries, has the potential to be twofold. Firstly, geographic representations of large data sets can be a powerful tool for imposing a first layer of interpretation on the data. We call this the value **of** geography. Secondly, information gathered from large volumes of text is relevant in order to answer a variety of traditional geographic research questions. This describes the value of digitized text **for** geography. The two roles of geography constitute the greater motivation for this thesis and are thus exemplified in the following two paragraphs.

The value of Geography. Geographic information and geographic representations in particular, can be seen as a prominent way to impose a first layer of information on large digital data sets. Two examples from different application fields are given by Crandall et al. (2009) and Andrienko et al. (2010). Crandall et al. (2009) map some 35 million images collected from Flickr for "revealing various interesting properties about popular cities and landmarks at a global scale" (p.761) (Figure 3). Andrienko et al. (2010) visually represent a global spatio-temporal data set on flue distribution, in order to detect particular characteristics.

Figure 3. Mapping Flickr images to Europe (altered from Crandall *et al.* 2009).

A third example will be given as a thought experiment. Imagine an extension of the above introduced Google Books Ngram Viewer where, additionally to the temporal plots, a map representation for a given topic is provided. This would clearly improve the semantic content of the retrieved information and allow for answering where the rise of computers in the 1950s was initially discussed and how it has since spread. However, this is a difficult challenge, since geographic information is not explicitly contained in written natural language. It is seamlessly embedded in the body of text, for instance in the form of place names (i.e. toponyms). Thus, geographic information has to be recognized and extracted before it can be used for further investigations.

The value for Geography. The information contained in large historic compilations of digitized books can be of vital importance for geographic investigations. Detailed information on how people describe their local environment, for instance, is crucial in applications such as *land use planning*, *natural resource management* or *crisis intervention*. However, the use of local information from written landscape descriptions does not only introduce new means for geographic applications, it also offers the potential for contributing to basic geographic research questions. One example is given by the *ethnophysiographic hypothesis*:

> *"People from different language groups/cultures have different ways of conceptualizing landscape, as evidenced by different terminology and ways of talking about and naming landscape features."* (Mark *et al.* 2007, p. 16)

Ethnophysiography aims at characterizing the basic way in which people perceive and describe the world. One prominent finding is that landscape concepts are subject to local variation. These local variations in concepts and terminologies are often referred to as *ambiguity* and *vagueness*, which are both uncertainties that we constantly deal with, and mostly successfully resolve, in our daily lives. However, they provide numerous challenges when we wish to represent and compare such information in a computer. Another aspect of the importance of local geographic information is covered by *naïve geographic knowledge*, as discussed by Egenhofer and Mark (1995). The authors argue that naïve geographic knowledge is crucial for bridging expert and lay people's concepts. They elaborate that "[t]oday's GIS do not sufficiently support common-sense reasoning; however, in order to make them useful for a wider range of people […] it will be necessary to incorporate people's concepts about space and time and to mimic human thinking" (Egenhofer and Mark 1995, p. 5). A more pragmatic take on the same issue is represented by White and Buscher (2012) from Microsoft research. They recently stated that local knowledge is key for knowing local interests, which has crucial "implications for search and recommendation systems" (p.1607).

Now, almost 20 years have passed since the introduction of the term naïve geographical knowledge and a decade since the first ethnophysiographic investigation was published (i.e. Mark and Turk 2003). But there are limited means for gathering local geographic information for large spatial extents. In this context, written landscape descriptions bear a great possibility to do so, as they are available for large temporal and spatial coverage. Accessing them might unveil particularities of local landscape concepts, which are only fragmentarily covered by state of the art empirical or ethnographic investigations. Recent developments in the field of *geographic information retrieval* (GIR) (Purves and Jones 2011) and the exploitation of user generated content (e.g. Goodchild 2007) might serve as a source of methods for extracting information from digital landscape descriptions.

Goal. The goal of this thesis is to use written landscape descriptions in order to unveil and investigate local landscape concepts. The thesis has two key objectives, each reflecting one of the above described roles, *of* and *for*, that geography has in the context of large compilations of digitized books:

1. We[1] aim at linking landscape descriptions to spatial footprints[2], reflecting the *value of geography* when working with large digital text data. This is mainly a methodological contribution, emphasizing the role of geography as a tool for analyzing large unstructured data.
2. We will extract local landscape information from large sets of landscape descriptions. This reflects the *value for geography*, in terms of contributing to the state of the art in fields such as ethnophysiography that aim to understand how people describe their environment.

1.1 Research Questions

The two contributions outlined in the introduction are investigated on the basis of one general and three detailed research questions. The general research question is:

> *How can vagueness and ambiguity present in unstructured descriptions of natural landscapes be captured such that geographic queries can be effectively resolved (for lay communities)?*

Through answering the general research question we aim to find ways of using descriptions of natural landscapes in order to retrieve information and resolve uncertainties inherent to geographic information in

[1] I decided to refer to the work and results in this thesis using the pronoun *we*, expressing that most of the work and decisions were influenced by collaborations with other people.
[2] In our context the spatial footprint is the spatial manifestation of a natural landscape description, which is the sum of all toponyms found in text and associated with geographic coordinates.

written language (i.e. *vagueness* and *ambiguity*). This new information is then applied in order to improve *geographic information retrieval* and to answer fundamental geographic research questions, such as from ethnophysiography.

Three detailed research questions are introduced in order to subdivide the general research question:

> **RQ 1:** *How can natural landscape descriptions be linked to space, with particular consideration of ambiguity in toponyms referring to natural features?*
>
> **RQ 2:** *How can local landscape concepts be captured from descriptions, under consideration of the vagueness associated with geographic concepts?*
>
> **RQ 3:** *Does the introduction of methods aiming to incorporate vagueness and ambiguity result in improvements in retrieval effectiveness for geographic information retrieval?*

These three questions will be recalled and answered at the end of this thesis in order to summarize and discuss all important findings. The next chapter of this thesis is on setting the scene, in terms of discussing relevant literature. From the presented literature we will resolve a set of research gaps that are closely related to the above research questions.

The structure of the work is reflected by the key objectives summarized above. Firstly, we will automatically ground toponyms from a large compilation of digitized landscape descriptions, in order to draw maps from text. This reflects the previously sketched role *of* geography. In a follow up investigation we will extend on this work by retrieving explicit local landscape information from these landscape descriptions. This information will be put into the context of state of the art ethnophysiographic work and will thus be used to contribute to the basic geographic research question on how people describe their local environment. This underlines the important role of the information in large corpora of written descriptions *for* the geographic domain.

Chapter 2 Setting the Scene

In this chapter we discuss a body of relevant literature, investigations and approaches, in order to set the scene for this thesis. The literature review will cover all central concepts and the theory needed for contributing to the general research question, as posed in the introduction.

The general research question contains complex and controversially discussed concepts and key words, such as *vagueness*, *ambiguity* or *lay communities*. These concepts, and others, will be discussed under the umbrella of two broad topics, *landscape research* and the *extraction of geographic information from descriptions*. Both these topics are discussed in individual chapters. Based on the findings from the body of literature, we will resolve a set of *research gaps*, which will then be used to introduce the *workflow* of investigations that frames this thesis.

Landscape Research. In the literature review we will, firstly, discuss different ways of conceptualizing landscapes, beginning with an etymological and philosophical point of view, which will then be broadened to empirical and ethnographic investigations. In this context, we will introduce the concept of vagueness, as it is associated with the indeterminacy of landscapes and landscape features[3]. Work on the conceptual definition of landscapes will thus be contrasted by work that aims at modeling and delineating natural landscapes, and landscape features in particular, from terrain data.

Extracting Geographic Information from Descriptions. Secondly, we will discuss literature on gathering geographic information from written descriptions. We will describe two recent initiatives. A first initiative, associated with information sciences, aims at automatically building a spatial index for text and thus provides us with the means for performing geographic information retrieval. The second initiative is less sophisticated in terms of the applied methodologies. However, it uses the mapping of text as a product, in order to perform follow up analysis of the semantic content of descriptions.

[3] We use the terms landscape features, geographic features and geographic objects as synonyms, even though Smith and Mark (2001) have shown that fine differences between these notions can have significant impact on the types of objects that fall into each category. In this thesis we use all three terms in order to refer to things that are often mentioned when people describe landscapes. Examples are mountains, hills, valleys and rivers.

2.1 Landscape Research

The notion of *landscape*, or *natural landscape*, has a long tradition in many different scientific fields, such as *landscape ecology* (e.g. Naveh and Lieberman 1984), *environmental psychology* (e.g. Gibson 1979) and *geography*. But where does the term *landscape* come from and what is its original meaning? Naveh & Lieberman (1984) argue that the probably earliest reference to landscape is contained in the *Book of Psalms* (48.2), where landscape is described as the sum of things, making up the beautiful view of *Jerusalem*, with its temples and castles. This early reference to landscape might be controversially discussed. Yet it reflects that landscape is an ancient concept used by people in many cultures for referring to the surrounding environment. "Human beings live in and experience landscapes, and they interpret and alter those landscapes through cognition and action" (Jett 2011, p. 327).

It might be true for many cultures that landscape is an archaic concept. However, in the light of recent research on cross-linguistic categorization it is reasonable to assume that there are exceptions to this rule (e.g. Burenhult and Levinson 2008). Thus, the "portion of the earth's surface that can be comprehended at a glance" (Jackson 1984, p. 8) that is ubiquitous in western languages and communication, is not naturally given. This was recognized and discussed in the foreword of the seminal book *Landscape and Language*, edited by Mark et al. (2011).

> "Perhaps the term 'Landscape' doesn't help here: according to the OED, it came into English at the end of the sixteenth century from Middle Dutch, hitch-hiking on the small easel paintings produced for the newly formed urban bourgeois market in such things. From there, it was rapidly generalized to views and vistas, and then more slowly to the Romantic landscape-appreciation and garden making of the eighteenth century. That sentiment born of a vanishing countryside is not the subject matter here. Instead the focus of this book is on our *Umwelt*, the terrain and water worlds we inhabit and exploit. As yet we have no better widely-accepted term, however, that captures this interdisciplinary domain." (Levinson 2011, p. IX)

We are challenged by a similar issue. In the following sections we will introduce manifold views on the term and concept of landscape, ranging from etymology to philosophy, psychology, geography and geomorphology. However, the original motivation for using the term landscape so prominently in this thesis is well captured in the above citation: We are in need of a theoretical fundament that explains characteristics of how people describe and conceptualize their surrounding environment, as for instance conveyed in written documents. This is a main pillar of this thesis.

2.1.1 Landscape Concepts

Etymology. The English term landscape has Indo-Germanic roots. *Land* is originally related to uncultivated-, and later changed to open-land (Kluge 2002). *Scape*, or rather skapi/skapja/skafti, is etymologically related to *creation* or *composition* (Müller 1977). Thus, etymologically speaking, landscape is an extent of the earth's surface, populated by *created* objects.

Renaissance of Landscape. Landscape became a prominent term in the renaissance, in particular in relation to renaissance paintings. The Dutch term *Landschap*, and the German term *Landschaft* were both used to describe natural scenery in 16th century paintings. Landscape painting was an aesthetic discourse with the environment which often lead to idealization (Simmel 1913).

Landscape in Geography. The introduction of landscape as a concept to geography is often associated with the work of Humbolt (Naveh and Lieberman 1984). Humbolt described landscape as the 'Totaleindruck einer Gegend', which is the holistic impression of a region (Hard 1970). Landscape, in the context of Humbolt, is often associated with a set of objects, landforms, and its aesthetic impression. Saur (1913) disagrees with this rather physical and distant view on landscape in such that landscape should be the primary object of study of geographers, or, *landscape is geography*. Saur suggests to study the morphology of the earth's surface but then, importantly, to apply it to "reveal the characteristics, traces, distributions and effectivity of human cultures […]" (Wylie 2009, p. 23). Jackson (1984), along the same line, argues that since humans play an active role in influencing landscape, the view on landscape is to be *democratized*. Accordingly, Jackson named the term *vernacular landscape* which exemplifies that the meaning of landscape can be beyond its physical manifestation, for instance by having a symbolic value transported by myths and local beliefs.

An early example of a perception based approach to define landscapes is described in Granö's seminal book *Reine Geographie* (Pure Geography) from 1929 (Granö 1997). *Reine Geographie* was an early attempt to bring the topic *landscape geography* to prominence. Granö originally called it *regional science* and announced it to be a "completely new version of geography as a field of research and teaching" (p.1). Granö links landscape to perception such that the meaning of a certain landscape is different in different contexts, which cannot be represented as a composition of physical features only. Thus, *subjective landscape* is a building block of regional science, meaning that landscape is a fundamental human concept in order to experience the environment on a meso-scale. The link between landscape and perception is also reflected in the notion that the distance between the observer and the observed has crucial impact. With increasing distance the phenomenal perception, of color, form or size, etc., changes. Thus, Granö identifies two scales of environmental perception, namely the *proximate* and the *landscape*.

The proximate environment is perceived with all five senses. By contrast, landscape is only visually perceived and only consists of earth and sky.

Tuan (1974) introduced *topophilia*. Topophilia is the love one establishes for his own home locality or "the affective bond between people and place or setting" (p.4). Emotional bonds reflect, but also effect, perception. Local people will always have a different perception of their locality than visitors, in such that a native usually has a more "complex attitude derived from his immersion in the totality of his environment", whereas a visitor's perception is mainly based on "using his eyes" (p. 63). Due to topophilia and emotional involvement, local people can sometimes struggle in correctly distinguishing tales or exaggerations which conflict with historical facts. This struggling can of course also relate to the necessity or choice to guard one's own *Weltanschauung* (i.e. world view).

Apart from theoretical or philosophical approaches, such as Granö's or Tuan's, perception is also an important element in official definitions, for instance, by the *European Council* where landscape is defined as "[a]n area, as perceived by people, whose character is the result of the action and interaction of natural and/or human factors" (§1). Another example is the *Historic Landscape Characterization* (HLC)[4] initiative. HLC was developed, for English Heritage and English local governments, for the purpose of emphasizing that landscape is mainly a product of perception (Fairclough 2006).

Perception and Cognition. For this thesis it is important to emphasize the important role of perception, mainly in making landscape an individual experience. Perception is initiated by the retina that scans grounds for different colors and brightness, in order to extract contours and symmetries, and to isolate individual figures (Hochberg 1978). Figures, at this stage, are individuals and highly dependent on the observer. In a follow up process figures are classified, such that they are identified as belonging to certain classes of objects. Usually, the first classification sorts objects into basic level categories (Rosch 1973). People are more likely to identify a figure as a chair, than as a long chaise or furniture. The extraction of figures from ground is mostly sensory-driven, whereas the identification of objects strongly depends on general knowledge and context (Marr 1982). The definition of perception overlaps with the concept of cognition, which in cognitive psychology is often described as the processing of information (e.g. Reitman 1965). Thus, cognition is clearly involved when figures are grouped into classes of objects, as it is described above. In this thesis we will use the term perception whenever we refer to the individuality of landscape descriptions. We will avoid the term cognition as it is usually associated with a broad range of psychological processes, such as *problem solving*, *learning* or *reasoning*, which clearly exceeds the focus of this thesis.

[4] HLC is a map-driven landscape classification based on historic information going back to the early 19th century.

The impact of individual perception will be further discussed when summarizing approaches that aim to investigate landscapes through empirical experiments and ethnographic investigations.

Summary Landscape Concepts:

- Etymologically, landscape is a composition of mainly created objects.
- Landscape became the central concept of an aesthetic examination of nature in 16th century painting.
- Perception is central to many theoretical enquiries on the relation between humans and the physical environment, often referred to as landscape. Perception is often used for explaining individual differences in landscape concepts.

2.1.2 What is Natural?

If children are asked to describe landscapes they usually list natural features, like lakes, mountains or hills (Volk and Steinhardt 2002), indicating that *natural* is more closely related to our understanding of landscapes, compared to attributes such as *urban* or *artificial*. A clear-cut answer to the question of what is a natural landscape is not of fundamental importance in the context of this thesis. However, since we frequently refer to natural landscapes, as opposed to artificial, populated or cultivated places, and since it is our aim to investigate the description of natural landscapes in written documents, in this section we will aim at giving a brief overview of the evolution of the term natural in a landscape context.

Before the 17th century landscapes were mainly perceived as either being cultivated or wilderness (Shaftesbury 1964). Wilderness, in German, is etymologically closely related to forest (Wald) implying that wilderness is not in arable use and thus uninhabited land (Zedler 1749). In this context it is not surprising that the term wilderness used to have a negative connotation. Wilderness is hostile and dangerous whereas cultivated landscapes are fundamental to life. In the 17th century nature becomes an aesthetic norm in England, indicated by the cultivation of extensive public parks and gardens. Thus, to the dichotomy of wilderness and cultivated landscape is added a new, intermediate concept, the *cultivated wilderness* or *nature*. This is a first indication of nature having a positive connotation. Nowadays, in German literature, wilderness is captured as "wild landscape, land in a natural state" (translated from Warhig 1994). The change in perception is best reflected by the name of the Swiss Alpine heritage organization: Mountain Wilderness Schweiz (Stremlow and Sidler 2002). Wilderness has become a subject of study and conservation such that, for instance, glaciers, prototypes of a wild mountain landscape, evolved from *montes horribiles* (Walter 1996) to *unique demonstration objects* in research in the course of only two centuries (Haeberli 2009).

In North America, and in particular California, John Muir is regarded as the "forerunner of modern environmentalism" (Worster 2008, p. 3). Muir, in 1917, argues that people have a passion for nature

derived from the "natural inherited wildness in our blood" (Muir 1917). This understanding of the aspiration of nature as an archaic human heritage is reflected in a contemporary political movement. Environmentalism in the early 20th century is regarded as a way to express the "emotional and material interdependence of humans and nature" and is part of the political program of the liberal democrats. Muir was not a politician himself but one of the first voices that demanded that the preservation of wilderness and the setting aside of national parks and wildlife sanctuaries is the responsibility of the government.

Vale (2002) in his seminal book *Fire, Native People and the Natural Landscapes* seeks a way to circumnavigate the simplicity of the duality of pristine and humanized landscapes. Vale's work is motivated by the abrupt change in how North America is described before the first European settlers arrived. Originally visualized as complete wilderness this vision was abruptly replaced by a concept of North America that was vastly human-modified by native people. Through a set of studies on the fire regime in North America before European settlement Vale shows that the duality of pristine and humanized landscapes cannot explain the mosaic of areas found, some intensively altered by native people and some dominated by natural process, but most of them somewhere in between the two extremes. Vale therefore suggests a hybrid, seven-part scheme that "spans the range of possible landscape conditions in the otherwise dichotomous distinction between "humanized" and "pristine" landscapes (p. 298). The scheme contains the landscape types *intensely-humanized*, *uneven-humanized*, *amplified-humanized*, *mosaic*, *natural*, *inhabited wilderness* and *untouched landscapes*. Natural landscapes are clearly on the pristine side of the range but still far from being untouched.

Summary Natural Landscapes:

- The connotation of natural changed over time. Wilderness was for instance considered inhuman before the 17th century and changed ever since to a valuable resource that is to be preserved from the impact of civilization.
- Nature has become a political agenda, reflected by environmentalism, introduced in the early 20th century.
- There is no clear-cut distinction between natural and artificial landscapes. Vale, in his landscape framework, introduced multiple grades of landscape states, examples are untouched, natural or inhabited landscapes.

2.1.3 Landscape Terms and Toponyms in Landscape Descriptions

Since landscape theories in the former section tend to underline the role of perception, we would like to emphasize different ways of how people describe landscapes. One research question pursued in a study on *Landscape and Language*, Mark et al. (2011), asks:

> "What is the denotational relation between landscape terms and place names?"

Landscapes have two representations in language. They are either described by terms, or represented as place names. Mark et al. suggest that these two representations are in a *denotational* relationship, which should be central to further investigations. *Denotation* is an expression borrowed from semantics and refers to the literal or objective meaning, which unambiguously translates from a sign (e.g. wording) to its meaning. In linguistics the denotation is sometimes referred to as the *dictionary definition* and contrasted by the *connotation*.

Landscape terms are generic classes of objects. Prominent examples of landscape terms are *mountains*, *hills*, *rivers* and *forests*. Landscape terms will be extensively discussed in the following sections. Here we will have a more detailed look at place names or toponyms[5].

Toponyms. Toponyms are proper names referring to individual landscapes or landscape terms, for example *New York* or *Mt. Everest*. Toponyms are considered to be one of the most important sub classes of proper nouns. Levinson (2011) has pointed out that toponyms and personal names are the only two domains of proper nouns with distinct processing areas in the brain. There is a debate in literature on the meaning, if any, of toponyms. Hollis & Valentine (2001), based on empirical investigations, argue that "[l]andmark names often contain a greater degree of meaning compared with people's names and country names that can be considered arbitrary" (p. 113). On the other hand, Coates (2006) suggests in his account on properhood that proper names are "a type of referring that discounts the sense of any lexical items (real or apparent) in the expression that is being used to do the referring" (p. 378). This is in one line of argumentation with Wittgenstein (1922), who states that "Der Name bedeutet den Gegenstand. Der Gegenstand ist seine Bedeutung […]"[6] (p. 203).

Both types of landscapes references, generic landscape terms and toponyms, can simultaneously be used for referring to the same landscape. The landscape around Zermatt, shown in Figure 4, could be described using the terms *mountain*, *Matterhorn*, *glacier*, *Hörnlihütte* and *Zermatt*, some of which represent generic descriptions (e.g. mountain, glacier) and others are specific toponyms (e.g. Matterhorn, Hörnlihütte).

[5] Toponyms are proper names of locations (e.g. New York, Mt. Everest or Golden Gate Bridge). We use toponyms as synonyms of place names or geographic references.
[6] "The name refers to the subject. The subject contains its meaning" (own translation).

Figure 4. The landscape of Zermatt, Switzerland. In the background the Matterhorn. (Source: Flickr, User: Craig McKerral)

Sara Shatford (1986) argued that "pictures are simultaneously generic and specific" (p.47). A picture of a bridge for instance refers to the generic object bridge and, simultaneously, to the particular bridge shown in the picture. Shatford's theory of generic and specific information goes back to Frege's *referential theory* (Frege 1994) on the meaning of language. In referential theory words have sense and reference. A particular reference, like the Golden Gate Bridge can have different senses (like connectivity, construction, power, etc.), and the sense of a bridge can have millions of references of which Golden Gate Bridge is one. Shatford uses the notion of sense and reference as a foundation to classify subjects of pictures. The sense of a picture, or its generic meaning, is called *Generic Of* (e.g. bridge). The *Specific Of*, on the other hand, refers to an individual object (e.g. Golden Gate Bridge). Shatford emphasizes the importance of providing both types of information when adding labels to pictures, in order to support simple (generic) and unambiguous (specific) identification of pictures in large collections. Shatford applies her theory to different types of subjects contained in pictures, such as persons, matter, time and space. Her theoretic framework is summarized in the following facet matrix (p.49, Table 1).

Table 1. The Panofsky-Shatford facet matrix.

	Specific Of	**Generic Of**	**About**
Who?	individual persons, animals, things	kinds of persons, animals or things	mythical beings manifested by objects
What?	individual events	actions, conditions	emotions, abstractions manifested by actions
Where?	individual locations	kind of place or feature	symbolized places
When?	linear time, dates	cyclical time, season	emotions manifested by time

Shatford's framework is frequently used in *information science*, and in *image retrieval* in particular (Goodrum 2000, e.g. Hollink *et al.* 2004, Laine-Hernandez and Westman 2006). An application of Shatford's theory of generic and specific descriptions of pictures to the domain of geography is reported in Edwardes and Purves (2007) where they seek to provide better access to collections of digital images through key words that reflect people's concept of place.

The following two sections are on investigations aimed at gathering information on generic descriptions of landscapes, mainly in terms of landscape features.

Summary Specific and Generic Landscape Descriptions:

- Landscapes are often described using generic landscape terms. Specific toponyms are used to refer to landscapes.
- Sara Shatford suggests that the location content of pictures and images has to be described using generic and specific information.

2.1.4 Empirical Investigations

In recent years a series of empirical investigations on the definition of natural landscapes were conducted with the aim of unveiling universals or category norms in the individual perception of landscapes, and landscape features in particular. Central to these investigations is the conceptualization of landscape as a whole consisting of parts (Naveh and Lieberman 1984). Tversky and Hemenway (1983) found evidence for such part-whole relationships by showing pictures of natural scenes to participants of an empirical investigation. By asking participants to list activities, parts and qualities which can be associated with the scenes shown in the photographs it turned out that 95% of all terms listed represented parts of landscapes rather than the landscape as a whole. Such parts, in a natural context, can be called *landscape features* (or landscape terms, geographic features or geographic objects). Landscape features are central in many empiric investigations aimed at defining basic levels. The term *basic level* was intensively investigated by Rosch (1978). Rosch argued that categories are inherent to human perception, in order to facilitate organizational schemes and that basic level instances in categories guarantee maximum information gain with only minimal cognitive effort, compared to *super-* or *sub-ordinates*. A well-known example considers the class *chair* as a basic level, *furniture* as a super-ordinate and *long chaise* as a sub-ordinate. However, it is not always straightforward to find such unambiguous examples of basic levels, and associated sub- or super-ordinates, respectively.

A number of efforts have explored the categorization of landscapes features. Battig and Montague (1969), for instance, conducted classroom experiments on category norms (later termed basic levels by Rosch) for 56 different categories. Of interest here, is the category *natural earth formation* where participants most commonly suggested the term *mountain*. Smith & Mark (2001) asked students to list *geographic objects*, *features*, *concepts* or *something that could be portrayed on a map*. Different phrasings accompanying the term *geographic* led to a divergence in the results. However, Smith and Mark's experiment was conducted in different European languages, as well as in North American ones, and results suggest similarities in terms of common basic levels, with, for instance, *mountain*, *river*, *lake*, *ocean*, and *sea* being prominent features for all formulations of the question.

Basic level landscape features are important building blocks of taxonomies for classifying the earth's surface into meaningful entities, attached to relevant labels. They represent a set of words which is assumed to be representative over large spatial extents and for different groups of people.

Summary Empirical Investigations:

- Landscapes are wholes consisting of parts.
- Categorization is inherent to perception. Its purpose is to organize information.
- Basic level categories are defined as a combination of maximum information content and minimum complexity.
- The identification of basic levels is complex. One reason is that basic levels highly depend on the context.
- Empirical investigations on the nature of landscapes have shown that sometimes the parts of landscapes can be considered basic levels. These investigations, however, often have a western focus.
- Basic levels of landscape parts are considered important building blocks for taxonomies, used to structure landscape knowledge.

2.1.5 Ethnographic Investigations

In contrast to basic level research, more recent work in *ethnophysiography* and *landscape ethnoecology* suggests that differences between landscape concepts might be very pronounced. On that account the ethnophysiographic hypothesis is: "People from different language groups/cultures have different ways of conceptualizing landscape, as evidenced by different terminology and ways of talking about and naming landscape features" (Mark *et al.* 2007, p. 16). Ethnophysiography and landscape ethnoecology both have a particular focus on landscape concepts and perceptions of local indigenous people. Both fields conduct ethnographic investigations, including field walks or interviews (Bohnemeyer *et al.* 2004).

Mark and Turk (2003) found that categories of convex landscape features and water bodies for Yindjibarndi people, indigenous to Australia, are fundamentally different from an official English-language gazetteer (AUSLIG) describing the same spatial extent. Individual Yindjibarndi terms for water

and convex geographic features are similar to English. However, at the basic level category the two languages are significantly different. For instance temporary and permanent water features are fundamentally differently conceptualized in Yindjibarndi language. Also, Yindjibarndi people do not distinguish between topography and spirituality, such that to comprehend Yindjibarndi geographic concepts, it is necessary to adopt a method of inquiry that allows treating the spiritual as real which conflicts with Western concepts of landscapes and landscape terms.

Maori people, as reported by Murton (2011), although using generic terms in place names that often can be translated into English, such as mountain, hill, ridge or plain, often reflect the history of their ancestors when naming landscape features. Since history remains the same, even if the location of a tribe changes, place names are simply relocated to the new environment and to *new* features. Therefore, Maori place names are often brought from their *homelands* and do not reflect individual or specific labels.

Navajo language uses toponyms to reference landscapes that are very similar to landscape descriptions. Toponyms for instance often contain the term *hoolyé*, which translates into *a-place-called*. A typical ending for toponyms is *–i* which is similar to a definite article and thus refers to a named entity. These two notions are the only way to distinguish landscape terms from toponyms in Navajo languages (Turk *et al.* 2011).

Burenhult and Levinson (2008) explore the outcome of nine investigations in different languages on existing landscape features and suggest that neither similar topography nor cultural models could explain variation in the use of categories. For instance the concave feature *valley* is not universally present in all nine languages. In *Marquesan,* valleys, rivers and villages share the same term. One potential explanation is that they often co-occur in space and that therefore one term is sufficient for representing all three features. In two languages there is no comparable term for valley. “This absence of terminology cannot be explained by an absence of the landscape feature in question […]” (p. 141). Burenhult and Levinson suggest that caution is required when applying the European concept of landscape to other cultures. There might be the universal fact that societies are aware of their environment. This, however, is not to be confused with the presence of the concept landscape.

An interesting case of naming and structuring the environment is reported in Heyes (2011) on the example of Inuit in Kangiqsualujjuaq. For individuals who have never been to Kangiqsualujjuaq the landscape must appear as being void, with an absence of distinct landscape features. However, “[t]he landforms and waters contain their creation stories, history, myths and ancestral legends […]. Cosmological ways of knowing the land and waters, which are not possessed by an uninitiated visitor, allow the Kangiqsualujjuamiut to communicate information about the environment among each other and

provide for them a way to anchor themselves in their surroundings" (p.191). There are for instance three names for the feature *hole in ice*; one that opens and closes with the tides, one which is bigger and doesn't close and one which is used by seals to breathe. The tides are a central element of the Kangiqsualujjuamiut landscape which is reflected by the nomenclature, which is adapted to its dynamic force of transforming landscapes.

The above mentioned cases are only a few examples of how fundamentally different landscapes can be conceptualized, depending for instance on local belief, culture or particular environmental settings. A more extensive discussion of particular ethnographic investigations on landscapes and landscape features are reported in Mark et al. (2011) and Johnson & Hunn (2010). In a nutshell, findings in both ethnophysiography and linguistics support the notion that people from different places and cultures use different categories to describe their environment (Turk *et al.* 2011). Mark et al. (2010) argue that a "naïve view of geographic categories implicitly asserts that categorizations are universal across all cultures, languages and landscapes." (p.41), which, of course, does not correspond with reality. Further, Mark et al. (2010) state that the variation in landscape concepts has implications on the interoperability of a Geographic Information System (GIS). The GIS-need for valid and interoperable geographic information is in conflict with the finding that people have different concepts of the same landscape. In order to guarantee the validity of geographic information each concept would need to be stored separately, whereas interoperability requires all these separate concepts to be interlinked. This is very difficult to facilitate, especially on a fine spatial granularity level.

The applicability of basic level landscape features in order to serve as building blocks of landscape taxonomies is shadowed by recent findings of ethnographic investigations. One particularity of both types of investigations, ethnophysiographic and empirical, is the focus on participants or local people and their responses, which causes sample sizes to be limited and investigations to only be representative for small spatial extents. On the other hand, the individual samples are rich in information and of fine spatial and semantic granularity.

Summary Ethnographic Investigations:

- Ethnographic investigations on landscape concepts have shown that there is significant variation in how people conceptualize their surrounding environment.
- This variation is in contrast with the aim of previous empirical investigations for resolving basic level landscape categories.
- Ethnographic investigations on landscape concepts are of great level of detail, and thus, often of limited spatial and temporal coverage.

2.1.6 Vagueness

The ethnophysiographic hypothesis (i.e. people have different concepts of landscapes and landscape features) can be associated with vagueness (Mark *et al.* 2010). On this account, vagueness is related to perception and categorization, such that the definition of a concept, for instance mountain, changes from observer to observer, driven by individual, group, gender, etc. variables or by context. Vagueness is not exclusively geographic, but also known to philosophy (e.g. Williamson 1996) and linguistics (Lakoff and Johnson 1980). Concepts used in everyday conversation are often vague, mainly to avoid needless complexity. A particular geographic view on vagueness is represented by the discussion of the nature of borders. For this thesis it is worth discussing both views on vagueness separately, however, it is important to note that there is no clear-cut distinction possible between linguistic and geographic vagueness since the semantics of a concept are often intertwined with its physical manifestation and vice versa.

2.1.6.1 Vagueness in Natural Language

Fisher (2000) argues that vagueness "is in our view and understanding of everything around us, and, most profoundly, embedded in our natural language" (p.7-8). Excluding vague concepts from everyday human language would hinder us from using most of our vocabulary. Interestingly, vagueness was used as "a dustbin category, into which one dumped any failure to meet the ideal of precision" (Williamson 1996, p. 70) until the end of the 1920s. In 1937 Max Black introduced the notion that "vagueness is positive" (Black 1937). Black argues that vagueness is an adaption to our need not to clutter up communication with irrelevant information or precision.

An often cited geographic example of a vague concept is *mountain* (e.g. Smith and Mark 2003, Fisher *et al.* 2004). Mountain is prominently used in everyday conversation, with for instance more than one billion web counts on Google[7]. We can use the concept *mountain* in order to describe a weekend as a "weekend in the mountains" or we can characterize a hike as "ascending a mountain". Both examples are informative statements, such that mountain augments both sentences with *comprehensible* information, which is understood and reproducible for most of us.

However, the concept mountain has no concise definition (e.g. Smith and Mark 2003), such that if humans weren't capable of dealing with vague statements, they would be forced to describe their weekend activities by using exact topographic parameters such as steepness, texture, curvature or specific coordinates. Firstly, this would be time consuming. Secondly, this would hinder us from using concepts that can be *perceived* and described in words and would force us to focus on *measurable characteristics*.

[7] gathered 27.06.2013

Vagueness, although easing communication, is very challenging when language has to be captured in formal systems, for instance in the context of machine translation (e.g. Chiang 2007). Williamson (1996), on this account, states that vagueness cannot be fully captured by formal languages and that the "[…] matter of vagueness get its urgency from sorites paradox" (p.72).

Sorites paradox is a cognitively intuitive way for introducing the fundamental inconsistency between formal logics and vagueness. Traditional formal systems rely on distinct boundaries that keep different concepts separate. As a consequence, state of the art formal systems capture a mountain as a crisp object. This is where the sorites paradox comes into play. One intuitive way of defining the region of a mountain would be to start at the summit, reflecting the assumption that the summit clearly belongs to the mountain. At the same time, however, it is obvious that a mountain is larger than just the exact location of its peak. Consequently, we also consider the surrounding of the summit as belonging to the mountain. More generally this means that if we are sure that a location belongs to the mountain we concluded that also its next neighbors must be considered as belonging to the same mountain. This is particular true if neighborhood is considered on a small spatial scale, such as 1 or 2 meters. The formal expression for this is:

$$\text{if } i \text{ (e.g. summit)} = \text{mountain} \rightarrow i_{+1} \text{ (i.e. surrounding)} = \text{mountain.}$$

By applying this rule iteratively we would classify the whole world as being one mountain. The paradox is that each iteration is only one or two meters distant from locations that were previously considered as mountain. However, one or two meters are certainly not enough in order to move from clearly mountain to clearly not mountain. In a nutshell: the sorites paradox is a test to see if a concept is vague in terms of not having crisp boundaries.

The problem of not having crisp boundaries is often approximated by multi-scale approaches (e.g. Wood 1996) or fuzzy logic (Zadeh 1965). Multi-scale approaches are discussed in the section on geomorphometry (§2.1.8). Fuzzy logic is basically an extension of the duality of being and not being by introducing a continuous degree of membership. However, "fuzzy logic has been explored in the analysis of vagueness in the early seventies by Lakoff (1973), but has been regarded as unsuitable for the analysis of language meaning […]" (Sauerland 2011, p. 185).

2.1.6.2 Vagueness of Landscape Features

We just showed that vagueness is intrinsic to many concepts in natural language. Another aspect of vagueness is related to the spatial manifestation of landscape features:

> *"Existing research on cognitive categories has standardly addressed entities on the sub-geographic scale: manipulable entities of the table-top world, objects of roughly human scale (birds, pets, toys) and other similar phenomena. For such entities, the 'what' and the 'where' are almost always independent. In the geographic world, in contrast, the 'what' and the 'where' are intimately intertwined."* (Smith and Mark 1998, p. 309)

Categorization is core to human perception (and cognition, as briefly discussed in §2.1.1) and categories are dependent on boundaries in order to be distinct (Rosch and Lloyd 1978). This is true for all sorts of concepts, however, in the case of landscape features boundaries are not only relevant for having distinct semantic definitions, but are also important in order to delineate concepts, or rather objects, in physical space. The delineation of boundaries of landscape features is surprisingly challenging. Consider for instance the following photograph and how complex it is to clearly distinguish individual features, such as mountains, hills or valleys (Figure 5).

Figure 5. Bird eye view of the Allgäu Alps.

The complexity of identifying individual landscape features is caused by the fact that they are commonly perceived as distinct objects which are attached to the continuum of the earth's surface (Smith and Mark 2003). "[L]andform features are often indistinct and features are not defined disjointly" (Dehn *et al.* 2001, p. 1008).

Smith and Mark (2003) argue that "[t]he kind mountain is not a product of natural selection, nor does it represent an artifactual kind with bona fide instances which have arisen as a reflection of special human intention or purpose." (p.412). They argue that landscape features are neither biological creatures, like ducks (*products of natural selection*), nor artificially built objects like cars. Biological creatures, as well

as artificial objects, have *bona fide boundaries*. Bona fide boundaries separate objects of different physical matter, like a duck from the pond. Objects defined through bona fide boundaries are called *bona fide objects*, a typical geographic example is islands[8].

Fiat boundaries, on the other hand, have two different formats. Either they are defined in a top down process, for instance by "drawing lines on a Map" (Smith 1995, p. 475), as in the cases of some country boundaries, for instance in North America or North Africa. On the other hand, boundaries of landscape features "are also at least partly of the fiat type, although here the boundaries may result from cognitive rather than from legal or political processes" (Smith and Mark 1998, p. 312).

Since this thesis is not on the elicitation of landscape features from continuous surfaces, we will not cover details on the topic "fiat parsing the elevation field", as for instance discussed in Ralph Straumann's thesis (Straumann 2010). The role of vagueness in our work is one of awareness. We are for instance aware of the different types of vagueness when gathering spatial and semantic information on landscape features from descriptions. One consequence of the presence of spatial and semantic vagueness is the need to choose a suitable methodology for representing and structuring landscape information. In the following two sections we will firstly discuss approaches for representing knowledge on landscape features. Secondly, we discuss work associated with geomorphometry that has a particular focus on the extraction and representation of landscape features from continuous land surface data.

Summary Vagueness:

- Vagueness is omnipresent in language and plays an important role in communication.
- Vagueness is the lack of precise definition.
- Landscape features are prone to linguistic and spatial vagueness.
- Linguistic vagueness is investigated in the course of ethnographic investigations on landscape perception.
- Spatial vagueness of landscape features refers to undetermined spatial boundaries (i.e. fiat objects).

2.1.7 Ontology of Landscape Features

Ontology in Philosophy. Smith (2003), in a book chapter on ontology and its meaning in philosophy, states that "[o]ntology seeks to provide a definitive and exhaustive classification of entities in all spheres of being" (p. 155). This view on the world is simplistic. The world is conceptualized as something that can be divided into individual entities or parts, which can then be grouped into classes, which are then all interrelated in one holistic classification. This definition of ontology stems from philosophy (i.e. Aristotle) and serves as a tool to answer questions such as "What classes of entities are needed for a

[8] Although the length of the boundary, or coastline, of an island can be considered a separate challenge (Mandelbrot 1967).

complete description and explanation of all the goings-on in the universe?" (Smith 2003, p. 155). Such questions are not compatible with vagueness. Entities, such as mountains, need distinct and consistent definitions in order to provide holistic classifications of all *goings-on in the universe*.

Ontology in Information Science. A more recent definition of ontology stems from *information science* and the *artificial intelligence* community, stating that "[a]n ontology is an explicit specification of a conceptualization" (Gruber 1993, p. 199). This infers that several conceptualizations of the same reality can coexist. The coexistence of different concepts could capture the variation of landscape concepts. Guarino (1998) classifies such coexisting ontologies as domain-, task-, or application-ontology, contrasted by the top-level ontology that aims at classifying "entities in all spheres of being" (i.e. the philosophical meaning of ontology). The reason why information science adopted ontology from philosophy is the *tower of babel* problem. Different knowledge-bases, from different domains, organizations or countries use different classification schemas and terminologies. Ontology is thus used to firstly apply a set of rules to organize information, which is then, in a second step, bridged across systems in order to guarantee interoperability (e.g. Smith and Mark 1998).

Specification is a central term in the information science point of view. Specification can have different meaning, varying from simply introducing taxonomies (i.e. taxonomy) to applying *descriptive logic* (Bittner and Winter 2004). An example of a taxonomy used to classify land cover in *Europe* is *CORINE* (§3.4.2). CORINE is motivated by the need for consistent and interoperable information on the state of the environment within and across member states of the *European Union* (including Switzerland). CORINE applies a hierarchical taxonomy consisting of three levels with 5, 15 and 44 sublevels respectively (Bossard *et al.* 2000). CORINE seeks to provide interoperable land cover data mainly by applying the same, clear-cut classification rules to the whole of Europe.

However, taxonomies often come at the cost of only representing expert classifications that do not reflect everyday concepts used by lay people. The CORINE taxonomy for instance suggests using the class *sparsely vegetated area* for large mountain landscapes. This is a sub class of *forests and semi-natural areas* and *open spaces with little or no vegetation*. This is clearly different from concepts of alpine landscapes represented in textual descriptions. Therefore, expert taxonomies are only of limited applicability to represent landscape concepts of local or lay people.

Formal Ontology. Formal ontology consists of a set of logical *axioms* and requires the information to be *complete* and *sound*. Completeness means that nothing relevant exists that is not stored in the ontology. Soundness is the requirement that there are no contradictions or redundancies in the information. If these two conditions are met, formal ontologies allow for *reasoning*. Reasoning is the *inference* of new

knowledge from existing information (Guarino 1998). If I am my mother's son and my mother is the daughter of her father, which all sounds reasonable, then the family relationship between my grandfather and me does not need to be stored explicitly, since it can be inferred by the application of axioms. Reasoning is a major motivation for designing a formal ontology.

Ontology and Geographic Information. There are a number of general frameworks aimed at applying (formal) ontologies to capture geographic information. An extensive discussion of potential applications of ontology in GIS is reported in Agrawal (2005). In the following section we will focus on a set of relevant examples. An early, or even first, discussion of using ontologies to capture *geographic kinds* was published by Smith and Mark (1998). They argue that the geographic domain, and geographic objects in particular, is different from everyday objects in terms of not representing table-top space. Geographic objects are tidily bound to the earth's surface "and this means that their spatial boundaries are in many cases the most salient features for categorization." (p.1). To this end, Smith and Mark argue that an ontology containing geographic objects needs to incorporate spatial relations such as topology and mereology. In the same seminal paper Smith and Mark argue that geographic objects and their delineation strongly depends on perception, which introduces inter-personal, inter-language, and inter-cultural variation and requires extensive experiments with human subjects. Such peculiarities of geographic information were the motivation for a detailed theoretical examination (e.g. Smith and Mark 2003) and for some of the above discussed empirical and ethnographic investigations (e.g. Smith and Mark 2001, Mark and Turk 2003). However, to this date there is still no implementation of an *ontology of geographic kinds* as it was originally suggested by the authors. One probable reason for this is the complexity introduced by variation, which contrasts with formal, sound and complete definitions for the geographic realm.

Recently, Kuhn (2011) elaborated on the use of ontology in the context of landscape and language, arguing that "[l]anguage studies could benefit to a much larger degree from computational approaches to knowledge representation and reasoning than they currently do" (p. 369). The central claim of the essay is to use ontology for specifying concepts, in order to guide interpretations, rather than to define the meaning of words. Thus, the question of eliciting the earth's surface into meaningful parts and attaching representative labels is circumnavigated by focusing on language use only, i.e. a geographic feature becomes a noun. The specification of nouns from different languages could then be compared or linked. The specifications could be represented using the DOLCE[9] ontology.

[9] www.loa.istc.cnr.it/DOLCE.html, visited 27.06.2013

We agree with Kuhn's line of argument and clearly share his view on using language for gathering information on landscape concepts. However, we will avoid using an ontology language for further specifications in a first attempt, since we primarily aim to shed light on the local variation of landscape concepts that are retrieved from landscape concepts as presented in large corpus data. Introducing specifications at an early stage would potentially have a smoothing or generalization effect on the retrieved information, which we wish to avoid. Additionally, we put a critical note on simplifying geographic features as being linguistic features. Theoretic enquiries on the nature of geographic features emphasize their bond to the earth's surface and the influence of its shape on perception. Detaching geographic features from their physical manifestation for us means ignoring an important defining element.

Ontology Framework for Geographic Information. There are a number of recent publications on ontological frameworks for geographic information, for instance from Bittner et al. (2009), Couclelis (2010) or Bateman et al. (2010). Bittner et al. (2009), as an extension of Bittner and Winter (2004), propose a spatio-temporal ontology that integrates geographic information. They suggest the use of a formal top-level ontology, dividing the world into *individuals* (e.g. Napoleon, New York City), *universals* (e.g. human, settlement), and *collections* (e.g. counties in New York State). These categories are self-identical through time, but associated with differing temporal properties. Thus, most properties of, and relationships between, these categories are time-dependent. The ontology is designed using a *Web Ontology Language* (OWL) based implementation of *Basic Formal Ontology* (BFO) (Bittner 2009), in order to allow automatic reasoning and soundness checks. Bittner et al. (2009) applied their framework by integrating two diverse land cover data sets, ARKIS and CORINE (§3.4.2). This use case recommends the use of TNEMO-S-U in order to consolidate existing taxonomies into one top-level ontology. However, this use case does not account for the complexity introduced by vagueness and the consequence that often no taxonomies or clear-cut definitions are available - not to mention that clear-cut definitions could already be considered a constrain for capturing natural variance in landscape concepts. Very recently, Bittner (2011) implicitly agreed with this notion by concluding that "[t]he need for geometric representations, many of which rely on relatively precise boundaries, conflicts with the need for sophisticated classification systems for scientific and integration purposes. Understanding the true nature of the problem may be a first step toward overcoming it." (p.848).

A user centered framework is presented by Couclelis (2010), where *geographic information constructs* are captured. The goal is thus not the representation of real-world objects, but a tide bound to user intentionality. A central term in this context is *objects of discourse* (Bibby and Shepherd 2000), which serve as the building block of the ontology. Everything which is used in discourse is an object of

discourse; examples are *New York*, *belief* or *Zeus*. Objects of discourse have four dimensions, the *formal*, *constitutive*, *agentive*, and *telic*. These dimensions describe specifications of objects such as *properties* (formal), *parts* (constitutive), *function* (agentive) and *purpose* (telic). All of these are prone to vagueness. Couclelis, however, only refers to vagueness in the agentive and telic context, concluding with: "different representations of phenomena must in principle be developed for different scientific or practical purposes" (p. 1792). Couclelis organizes the four dimensions in a hierarchical schema, representing increasing *semantic content*. Semantic content reflects the demand of cognitive capabilities from the *decoder*, such as awareness, perception or intentionality. The process of *semantic contraction* is then defined as the stepwise draining of semantic information from objects of discourse, beginning with information that requires the most sophisticated capabilities to be recognized (i.e. intentionality). Thus, "the hierarchy generated by means of the semantic contraction procedure is characterized by well-defined semantic and logical relations between levels. This facilitates understanding how heterogeneous geographic entity representations may stand relative to one another" (p. 1805). This framework has not been applied to data so far. Its usability for storing and structuring geographic information has first to be proven. We would expect that vagueness affects definitions of objects of discourse not only on the levels with most semantic content, but even on the coarse semantic levels, where only properties of single objects are defined. Derungs and Purves (2007), for instance, indicated that people fundamentally disagree on the threshold height of a mountain. Duce and Janowicz (2010), among other things, discuss variation of river concepts and the special case of rivers in Spain, where they are dry for most of the year. Such fundamental disagreement between concepts on the property level could be a critical issue for the applicability of Couclelis's ontology. If concepts disagree on the property level, they are to be stored as completely different objects of discourse in the ontology. This can have critical impact on the size and the general applicability of Couclelis's framework.

Bateman et al. (2010) report on a *linguistic ontology of space for natural language processing* where they "present a detailed semantics for linguistic spatial expressions […]" (p. 1). Bateman et al. (2010) argue that their formal ontology, implemented as an extension to the *Generalized Upper Model* (GUM), covering all particularities of *SpatialML* (Mani *et al.* 2008), and specified using *OWL*, accounts for the flexible relationship between spatial language and its context dependent interpretations. Bateman et al. (2010) use a two-level architecture, where on the first level the linguistic ontology provides the semantics of the *raw* terms, such as left/right. On the second level, the raw terms are applied to spatial interpretations. Bateman et al. (2010) argue that this way, vagueness of spatial relations, such as in the notion *proximity*, can be captured in empirical investigations and separately added to the second level of the ontology, without having an impact on the first level. However, a spatial relation such as *proximate*

might require many different contextual interpretations. In fact, the problem of vagueness of geographic concepts and its impact on the applicability of ontologies is not solved by Bateman et al. but shifted to a peripheral level. The consequences are the same, the second level of the ontology, where each notion is associated with an interpretation will be of immense size and complexity.

Domain Ontology for Geographic Information. A well-known application of a domain ontology gathered from natural language documents is reported in Kuhn (2001). The ontology aims to describe human activity and focuses on the German traffic code system, gathered from the official handbooks. The traffic code description is used to deduce a formal cross tabulation of *actions* (verbs) and *objects* (nouns) that *afford* the action. These action-object relations are then hierarchically ordered using *lexical entailment* of verbs, as introduced by Fellbaum (1998), which helps achieve "a layering of the actions in the domain of car driving according to the German traffic code" (Kuhn 2001, p. 626). Kuhn argues that the activity centered domain ontology helps to structure the complexity of human conceptualizations of the environment, which he believes is a consequence of the growing complexity of human activities. With the traffic code system Kuhn selected a relatively well structured domain. Traffic rules are necessarily unambiguous and as precise as possible, since confusions lead to legal implications. Other domains, such as natural landscape descriptions are less *organized*. The application of Kuhn's approach to such an unconstrained domain might be very challenging.

Duce and Janowicz (2010) argue that the formalization of landscape concepts is the core of semantic interoperability. On the other hand, the authors emphasize that standardization, i.e. the agreement on a shared conceptualization, often means losing local variation. Their take on that is to introduce *microtheories*, such that each country is represented by a specific land cover ontology, formalized using OWL. The definitions of feature types for the individual microtheories are gathered from natural language use. All microtheories then contribute to one global land cover ontology by computing the *least common subsumer*. It appears to be a reasonable take to represent individual ontologies on local scale and to use these for generating a global onotology, however, this still leaves us with some uncertainties or vague definitions, for instance:

- *What if local scales are not suitably represented by countries?* Consider for instance Switzerland where we have four different language groups on a relatively small spatial extent.
- *What happens with local features?* Underlying the approach is the assumption that all features exist on all scales but are differently represented. However, it is feasible that features disappear or completely change their meaning, such that the different definitions cannot be aggregated.

The above examples of formal ontology application represent top down information, i.e. experts define a taxonomy and all associated rules and roles. Thus, for the sake of semantic interoperability and reasoning, these approaches often come at the cost of over-specification, which can conflict with lay people's concepts and which lacks flexibility for capturing natural variability.

2.1.7.1 Folksonomy

Some recent approaches to organizing information take on a different approach. They try to deduce structure from user generated content, such as tags in descriptions of social media contents. We subsumed a set of respective approaches under the umbrella of *folksonomy* and discuss it here. It is common to all these approaches that the opinion of relatively few experts is replaced by the participation of large numbers of users, or lay people.

"The word 'folksonomy' is a blend of the words 'taxonomy' and 'folk', and stands for conceptual structures created by people" (Hotho *et al.* 2006, p. 411). The major difference between folksonomy and ontology is the folk focus of the former. Folksonomies almost exclusively represent bottom up classifications of lay people, often stemming from tags, used on social media platforms (Vander Wal 2007). Egenhofer and Mark (1995), in a geographical context, refer to such information as *naïve geographical knowledge*. They argue that such knowledge is central in improving our understanding of how people describe the world in everyday encounters and to developing systems which are capable of being used without recourse to more formal models of space.

Tags are the building blocks of folksonomies. The rise of social media applications, such as del.icio.us, youtube.com, flickr.com or facebook.com, and the introduction of *tagging*[10], lead to the production of extensive amounts of tags in a relatively short time (Hotho *et al.* 2006) (e.g. Figure 6). Figure 6 is an example of a photograph uploaded to flicker and described by seven tags. The tags reflect geographic locations, such as *Switzerland*, *Alps*, *Pennine Alps*, *Zermatt* and *Matterhorn*, and activities, such as *vacation* or *trips*. Flickr contains some 200 million georeferenced and tagged pictures from all over the world.

[10] As *tagging* we consider the use of a set of words to describe media uploaded to a social media platform.

Figure 6. An example photograph uploaded by a user to flickr and described using some tags (Source: Flickr, User: Craig Stanfill)

There is a large body of literature referring to applications of folksonomy. Hortho et al. (2006) and Jäschke et al. (2007) both have an information retrieval focus on folksonomy. Hotho et al. (2006) present *FolkRank*, a graph based model, to rank and search tags and thus exploit the structure of the del.icio.us folksonomy. Jäschke et al. (2007) suggest use of folksonomies to recommend tags to users, thus simplifying the tagging process. They compare the above mentioned FolkRank algorithm with a user-based approach for computing similarities, for instance between users.

Gruber (2007a) and Chen et al. (2010) discuss the use of folksonomy in an ontology, or rather knowledge representation, context. Gruber (2007a) argues that technologies of the *Semantic Web*[11], which are best represented through ontologies, are to be applied to data of the *social web*. Gruber calls this mash-up *ontology of folksonomies*. The working group aiming at generating this social web ontology, namely *tagcommons.org*, has not been very active since 2007. However, the general idea of applying formal specifications from structured data and reasoning to data of the social web is still present in literature. Chen et al. (2010) stated that previous approaches of generating ontologies from folksonomies did not successfully take human thinking into consideration. Thus, Chen et al. resort to the theory of basic levels from *cognitive psychology* (e.g. Rosch and Lloyd 1978). The recognition of basic level categories from tags is implemented as a tag-clustering challenge, incorporating *tf-idf values*[12] (Equation 1) of tags. Basic level categories are then used to populate an ontology, which is optimized to represent bottom up user concepts.

[11] www.w3.org/standards/semanticweb/ visited 08.06.2013

[12] Tf-idf, i.e. term frequency – inverse document frequency, is a standard measure in information retrieval for normalizing the frequency of term occurrences in a document with the expected frequencies gathered from a large compilation of documents. We will frequently compute tf-idf values in this thesis. A comparable measure that became state of the art in information retrieval is *BM25*. Compared to tf-idf, BM25 does also incorporate text length. This is not necessary for our work, since most texts have comparable length.

Equation 1. Term frequency-inverse document frequency (tf-idf), with *tf* being the number of occurrences of a term *t* in a document, *N* being the total number of documents and *n* being the number of documents containing the term *t*.

$$tf - idf = \frac{\mathrm{tf}}{\log \frac{\mathrm{N}}{\mathrm{n}}}$$

Folksonomy in Geography. We are not aware of an explicit use of the term folksonomy in the realm of geography, or GIS in particular. Nevertheless, there are many approaches that implicitly relate to folksonomies, seen more often since tags have been used in order to deduce descriptions. Hollenstein and Purves (2010), for instance, use georeferenced photographs from Flickr to compute delineations of city cores. They computed spatial densities from tags, such as *citycenter*, *downtown*, *central* or *innercity*. Gschwend and Purves (2012), among other things, mapped the distribution of a set of geographic categories, resolved from previous empirical investigations (i.e. Purves *et al.* 2011). As a data source they used Flickr photographs and photographs with longer text descriptions from *geograph.org.uk*, both information sets are georeferenced. Wing and Baldridge (2011) introduced *Textgrounder*, an application to georeference Flickr photographs to geographic space. As a preprocessing step they compute tags that are particular for cells of a continuous grid, covering the earth's surface with a maximum resolution of 10km. This tag populated grid could be considered a place related folksonomy if the content would be used for further geographic analysis.

The defining element of folksonomy is the use of user generated content. Everything that uses social media tags to infer information could be considered a folksonomy. There are no other conditions to meet in order to consider something a folksonomy - in contrast to the methodological paradigm of using descriptive logic in formal ontology. In the context of this thesis, namely landscape analysis from text descriptions, we will relate to the folksonomy theory mainly as a bottom up approach for gathering landscape concepts. The use of plain text instead of tags could be considered a novelty.

In the following section we will briefly review literature with the aim of retrieving landscape information from morphometry. We consider this a complement to the above described approaches for structuring semantic information of landscapes. After introducing means for gathering and structuring information on the *what* component of geographic features, we will now focus on the *where* component, in terms of the spatial manifestation of landscapes.

Summary Ontology and Folksonomy:

- Formal ontologies allow inference of new knowledge (reasoning) at the cost of requiring complete and sound information.
- There are numerous frameworks that propose upper-level geographic ontologies but only few implementations. Most implemented geographic ontologies only cover particular domains (i.e. domain

ontologies).

- Using ontologies for structuring geographic information is challenging, since human concepts often vary (linguistic vagueness) and since most landscape features have undetermined boundaries (spatial vagueness). Additionally, formal ontologies are often based on expert taxonomies. These taxonomies do not usually overlap with lay people's concepts of their environment.
- Folksonomies are often considered as counterparts to ontologies. They are a loose concept, mainly defined by the type of input data (often user generated content) and the data structure being based on the opinion of often a large number of contributors.
- GIS know only few explicit, but a sizable number of implicit, applications of folksonomy for representing geographic information.

2.1.8 Geomorphometric Investigations of Landscape Features

We emphasized that linguistics, philosophy or geography know many approaches aiming to define landscapes and landscape features by conducting empirical investigations. In the previous section we discussed knowledge structures to formalize landscape information. However, most of these approaches only shed light on the *what* aspects of features, for instance by uncovering properties or associations.

Consequently, this section will have a focus on approaches that aims to define the *where* perspective of landscape features. This includes the modeling, the delineation and the locating of features. These approaches, usually associated with *geomorphometry*, share the notion of landscape features as being bound to the earth's surface (Smith and Mark 2003). Geomorphometry is defined as the extraction of land surface parameters and objects from *Digital Elevation Models* (Pike *et al.* 2009). This definition implies the broad focus of geomorphometry and the potential role it plays in different scientific disciplines such as *hydrology*, *geomorphology* and *glaciology*. In this chapter we will focus on a subset of approaches, aiming at describing and extracting landscape features, such as mountains or valleys, from the continuous elevation field.

Taxonomy of Approaches. A range of surface parameters is used to characterize and quantify land surface. *Slope* and *aspect* are calculated as the first derivatives of the elevation, *curvature* is a second order derivative (e.g. Kienzle 2004). These surface parameters are calculated using *focal moving windows*, where the window size (e.g. 3x3 raster cells) has crucial impact on the results (Wood 1996). Surface parameters can be combined to compound indexes. Examples are *topographic wetness index* (e.g. Beven and Kirkby 1979), *stream power index* (e.g. Moore *et al.* 2006), and *geomorphologic classifications.*

Geomorphologic Classifications. Wood (1996) distinguishes geomorphologic classifications into one group that classifies the surface into homogeneous regions and another group that identifies individual landscape features. Geomorphologic classifications of the former type associate each grid cell of the

elevation model with one landform class. Such classifications can be steered by a priori knowledge, i.e. supervised, or they can be unsupervised. A priori knowledge usually has the form of threshold values of surface parameters, training data or information taken from literature (e.g. Wood 1996). Unsupervised approaches for classifying the landscape into landscape features usually derive boundary conditions of surface parameters from global measurements (e.g. Deng 2007).

Unsupervised and Supervised Landform Classifications. Iwahashi and Pike (2007) report on an unsupervised geomorphologic classification compound of the surface parameters slope, curvature and texture, i.e. the number of local maxima per area unit. Iwahashi and Pike group the land surface into a maximum of 16 landform classes. The result represents the terrain as a patchwork of landform values (Figure 7, left). Wood (1996) reports on a supervised landform classification algorithm which outputs a set of landform objects. Possible landform objects are *peaks*, *pits*, *channels*, *ridges*, *passes* and *planes*. The classification incorporates measurements on different scales. Window sizes and the minimum drop, which is used to grow local maxima to summit regions, are provided as input parameters to the algorithm (Figure 7, right).

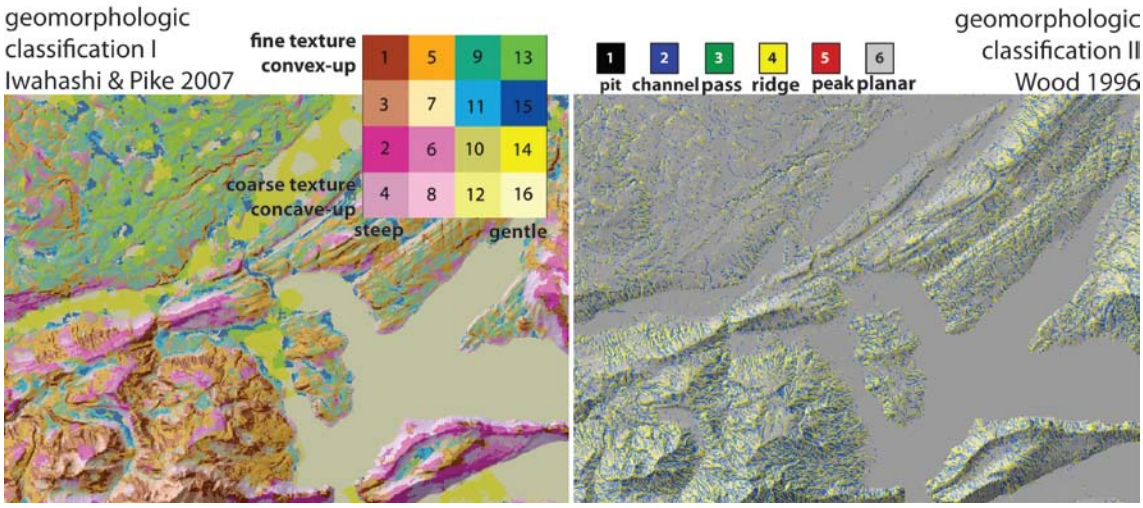


Figure 7. Geomorphologic classifications of the Digital Elevation Model in the region of Lucern.

Both classifications output information that has to be translated into meaningful concepts before it can be linked to human descriptions of landscapes. Thus, for instance, the class 7 of Iwahashi and Pike does not exist in natural language and also Wood aimed at using labels for his classes that are not directly associated with landscape features, such that some types of *channel* could be called *valley* in natural language, or *pit* could become the *summit* or the *mountain*.

Landscape Feature Extraction. Some examples of geomorphologic classifications that delineate landscape features are reported in Fisher et al. (2004), Straumann and Purves (2008) and Sinha and Mark (2010). Fisher et al. (2004) set out a multi-scale approach to perform a fuzzy classification of *peakness* to answer the question *Where is a mountain?* Straumann and Purves (2008) describe a region growing algorithm to identify valley floors, seeded by *thalwegs* and constrained by a threshold gradient. Straumann and Korup (2009) used these valley floors to successfully quantify postglacial sediment storage at the mountain-belt scale. Sinha and Mark (2010) calculate *topographic eminence* in terms of "landscapes that rise up conspicuously from the ground to visibly dominate the landscape […]" (p. 105). A simple manipulation of the parameters for *relative peak height* and *distance* can thus be used to establish alternative conceptualizations of eminences for the same landscape.

Algorithms for extracting individual landscape features clearly demonstrate the limitations of physical models for parsing the earth's surface. While there are numerous approaches for extracting geomorphologic and hydrologic features, such as channel networks (e.g. Tarboton *et al.* 1991) or catchment areas (e.g. Freeman 1991), there are relatively few examples where features that are prominently represented in communication, such as mountains or valleys, are extracted (except from the ones we discussed above). However, ethnophysiographic work suggests that around 100 natural features are needed to represent an individual landscape concept[13]. 100 is a large number considering the work required for extracting one feature type in suitable means. Additionally, the variation of landscape concepts is assumed to take place on reasonably small scales, such that even in a country the size of Switzerland definitions of individual features could be prone to variation. This has crucial impact on approaches aiming to physically model individual features, since each variation of a concept has to be considered separately – which is reflected in the approach of Sinha and Mark (2010) discussed above.

Linking Geomorphometric and Semantic Information. Derungs and Purves (2007) and Gschwend and Purves (2012) link semantic information of landscape features with information from geomorphometric classifications. Derungs and Purves (2007) used a questionnaire to conduct an empirical investigation on a set of surface parameters, such as *slope*, *elevation* and *dominance/prominence*, and evaluated the individual impacts mountain-perception in pictures. Results suggest that there is only limited inter-subject agreement with, for instance, a standard deviation of a threshold height of mountains of 700m (this is around 25% of the physically possible variation). The results from the questionnaire were then used to model *cognitive mountains* in Switzerland. This approach is again of limited applicability to the range of feature types that is used to describe Switzerland. Gschwend and Purves (2012) classify tags used to

[13] The number *100* stems from discussions with Prof. David Mark.

describe georeferenced pictures by the use of geomorphometric information and thus show how descriptions change with changing topography.

Summary Geomorphometry:

- Geomorphometry is defined by the extraction of land surface parameters and objects from elevation models. Land surface parameters can then be used to design compound indexes or geomorphologic classifications.
- Geomorphologic classifications can either be supervised (e.g. Wood 1996) or unsupervised (e.g. Iwahashi and Pike 2007).
- Often the class labels used in geomorphologic classifications do not correspond with terms used in communication. They first need to be translated in order to be meaningful.
- The work required for extracting an individual feature type from an elevation model is in an imbalance with the large number of feature types that manifest one landscape concept. Additionally, feature extraction approaches usually ignore that landscape concepts can undergo significant variation even on local scales.

2.1.9 Summary

The reason for having a closer look at literature on landscape relevant research is that characteristics of descriptions and conceptualizations of landscape are a central topic of this thesis. We are particularly interested in how people describe landscapes in written documents and if such descriptions could be used to deduce landscape concepts. Landscape relevant research as presented in the previous chapter is divided into, firstly, theoretical frameworks, secondly, investigations of landscape concepts, and landscape features in particular and, thirdly, ways of gathering geomorphometric information on landscape features.

Theoretical Landscape Frameworks. *Landscape* is an **ancient** concept used to refer to the surrounding environment and is thus central to **experiencing the world in everyday encounters**. Landscapes are seen as **wholes, consisting of parts**, i.e. landscape features. Landscape features are perceived as objects, associated with attributes. These objects, however, are bound to the earth's surface and are thus often characterized through their shape. Geographic theories on the nature of landscapes emphasize the crucial role of **perception,** which turns landscapes into individual experiences. This is for instance reflected by the notion *natural* in **natural landscape**, which underwent significant changes over time. *Natural* was originally considered as being dangerous, but nowadays its perception has changed to being a quality to be protected.

Investigations of Landscape Concepts. Investigations of landscape concepts aim at retrieving information on the existence, the prominence and the definition of individual **landscape features** in different cultures or languages. Two types of landscape investigations can be identified. Firstly, a series of **empirical investigations** on geographic features that aim at finding universal category norms, i.e. **basic level** categories, by conducting classroom experiments. A second type of investigation, usually applying

ethnographic research methods, aims at describing **local (indigenous) landscape concepts**. Ethnographic investigations contrast with the aim of finding basic levels, such that they emphasize **significant local variation in landscape concepts**. This local variation is often associated with **vagueness**. Vagueness is most often successfully resolved in everyday conversation. However, it constitutes a considerable challenge if landscape information is to be stored in a computer. **Ontologies** are state of the art knowledge structures in information science, however, the presence of vagueness in geographic information, and its implications on the sound and completeness heuristics, must be considered a crucial limitation. More recent applications for representing individual concepts (on all sorts of things, such as music, images or books) suggest the use of **folksonomy**. Folksonomies are often described as informal knowledge structures, gathered from tags of social media content and can thus be considered to represent bottom up or, in our case, **naïve, geographical knowledge**.

Geomorphometric Information. Empirical and ethnographic investigations on landscape concepts were introduced as a means for gathering information on landscapes and landscape features and describing its properties and constellations. **Geomorphometry**, on the other hand, can be seen as a set of tools for retrieving information on the **physical manifestation** of landscape features. Geomorphometric information can either be used for **characterizing locations**, for instance by computing surface parameters or **geomorphometric classifications**, or the geomorphometric information is used for **extracting individual landscape features** from the continuous earth's surface.

The resolution of research gaps from the presented body of literature on landscapes will follow after discussing research on *extracting geographic information from descriptions*. In a nutshell, the discussion of landscape relevant research has equipped us with important theoretical foundations that sometimes have considerable practical implications (e.g. vagueness). In the following section we will mainly focus on methodologies, rather than theories, that allow the retrieval of geographic information from text descriptions.

2.2 Extraction of Geographic Information from Descriptions

The frequently evoked data avalanche (e.g. Miller 2010) has long since reached human and social sciences. Michel et al. (2011), for instance, report on a quantitative investigation of cultural trends based on some 5 million digitized books. Thus, investigations in human and social sciences will increasingly

incorporate digital or digitized text and automatic data processing – simply due to the fact that not considering this data would mean to ignore available information. However, automatic data processing requires new approaches to extract information from unstructured text, as well as a critical way of dealing with biases that can occur in all steps of the processing chain, where natural language is converted into machine readable bits and bytes and then, into numeric representations (Boyd and Crawford 2011).

Bodenhamer et al. (2010) shed light on the role of Geography in this context by stating that "[s]cholars now have the tools to link quantitative, qualitative, and image data and to view them simultaneously and in relationship with each other in the space where they occur." (p. ix). Geographic information added to digitized text, from this perspective, allows the detection of spatial patterns, relations, or changes in time. However, this is accompanied by one key limitation and one consequence.

The limitation is that geographic information is usually not explicitly available from text. We need automatic means of extracting geographic information, such that it can be used to conduct further geographic investigations, for instance on the semantic content of descriptions. The consequence, on the other hand, is that the mentioned risk of not being critical in dealing with human sourced data particularly applies to geography. It is well known that the representation and intersection of spatial data affords decisions that influence the results, and it leaves room for diverse and sometimes contradictory interpretations (e.g. Why space is special? O'Sullivan and Unwin 2003). This obscures the *real* meaning of the data and demands critical approaches.

In order to discuss the limitations and consequences in connection with mapping text, we will mainly focus on two fields of research in the following chapter. Firstly, we discuss literature and methodologies associated with *geographic information retrieval* (GIR). GIR is closely related to information sciences and aims to resolve spatial footprints from text, in order to allow the retrieval of spatially relevant information (i.e. GIR). GIR has sophisticated means for automatically parsing text for occurring place names. We will discuss a set of GIR-related topics, such as *toponym ambiguity and toponym disambiguation* and *disambiguation of natural features*. However, GIR often omits to conduct further investigations, for instance using spatial footprints, in order to explain the semantic content of a description. Additionally, state of the art GIR often applies *bag of word* approaches (e.g. Manning *et al.* 2008), which are prone to change or ignore the context of the information. Thus, we will discuss a second body of literature associated with the topics *literary GIS* and *critical GIS*.

2.2.1 Geographic Information Retrieval

From IR to GIR. The retrieval of geographic information from unstructured sources is often associated with *Geographic Information Retrieval* (GIR). GIR is a combination of methodologies from GIS and *Information Retrieval* (IR). Larson (2011) defines IR as being "concerned with storage, organization, and searching of collections of information" (p.15). The medium of information is usually an unstructured source of information, such as text documents, images or videos. However, the extraction of the right piece of information from the right unstructured source is "not a simple task, and involves not only the technical aspects of constructing a system to perform such selection, but also aspects of psychology and user behavior […]" (Larson 2011, p. 15). Psychological considerations on particularities of spatial information, namely information on landscapes, were established in the previous chapters.

Is Geographic Information Particular? In this chapter we pay particular attention to methodological aspects of GIR. "It is only in recent years that much attention has been paid to the development of computer systems to retrieve geographically specific information from the relatively unstructured but immense resource of documents […]." (Jones and Purves 2008). Furthermore, Jones and Purves (2008) argue that classical approaches from IR, i.e. string based indexing and search, lack some of the specifications of geographic information, such as spatial qualifiers, toponym ambiguity, geographic relevance, spatial autocorrelation or geographic query expansion. Thus, classical information retrieval is successful in retrieving relevant information on queries describing relatively simple spatial settings, such as *What is the highest mountain in Africa?,* where the spatial compartment *Africa*, as well as the spatial preposition *in* can both be resolved by retrieving documents that contain the two strings *highest mountain* and *Africa*. By contrast, queries on complex spatial constellations, such as *What mountains can be seen from top of Breithorn?* are usually poorly resolved by classical information retrieval. A correct answer must incorporate geographic fundamentals such as *What is a mountain? How can visibility be modeled?* and *Which Breithorn is meant by the query?* Such background knowledge cannot be approximated by treating text as a *bag of words*[14] (e.g. Chowdhury 2010).

GIR Systems. There are numerous implementations of GIR Systems. GIR Systems are architectures that allow queries to be prompted, including spatial and sometimes temporal dimensions, and to retrieve result sets and spatial representations. Examples of GIR Systems are GIPSY (Woodruff and Plaunt 1994), SPIRIT (Purves *et al.* 2007) and STEWARD (Lieberman *et al.* 2007). An extensive list of GIR Systems is

[14] *Bag of words* is a metaphor, emphasizing that text is considered a set of words, where only the wording is used as a source of information. Linguistic information, such as word order, sentence structure, grammar rules or syntax is usually ignored in a bag of words approach.

summarized and discussed in Palacio et al. (Palacio *et al.* 2010, p. 96, Table 2). In the following sections we will look under the hood of GIR Systems, by discussing their major tasks and components.

Components and Tasks. Jones and Purves (2008) (and Purves and Jones 2011) recognized the following list of tasks as being of relevance in a GIR System:

- *resolution of geographic references*, toponym locations, from unstructured text
- interpretation of *vague and vernacular place names*
- *geographic indexing* of document footprints
- *geographic relevance ranking* of document footprints for spatial queries
- effective *user interfaces*
- methods for *evaluation*

Not all of these tasks are relevant in the context of this thesis. Of major importance is the resolution of geographic references from unstructured text, which will be discussed in a separate chapter. In the following paragraphs we will focus on the three tasks, *indexing*, *ranking* and *evaluation*, which we all consider as being important for this thesis.

Spatial Index. Indexing techniques that use the words contained in documents, e.g. string index, are well established methods in IR. Usually documents are converted into an *inverted file structure*, i.e. a list of words associated with lists of documents that contain this word. Spatial indexes, on the other hand, are used to allow spatial information retrieval, such that information can be retrieved which is linked to a certain region of interest, or: "In order to handle spatial data efficiently, as required in computer aided design and geo-data applications, a database system needs an index mechanism that will help it retrieve data items quickly according to their spatial locations" (Guttman 1984, p. 47).

In the case of a natural language document, the spatial index is computed from the spatial footprints of documents. Document footprints can be of different formats, such as one or several points per document, lines, bounding boxes, convex hulls or density maps (e.g. Vaid *et al.* 2005). In the simplest case, each document is represented by a single point, which can then for instance be used to build a *quadtree index*, where space is tessellated into quadrants of different resolution, depending on the local point density (e.g. Samet 2006).

A particular challenge is introduced when different types of indexes are used in combination, e.g. string, spatial and temporal indexes for the same data source. Palacio et al. (2010) report on an investigation where they combined all three types of indexes. A more extensive discussion of this work will be covered in the following section on evaluation.

Spatial Ranking. Ranking is the process of transforming a query into a ranked list of documents – usually by using indexes. Text ranking is usually processed by incorporating relative frequencies of query terms within documents, compared to frequencies in the whole corpus (e.g. tf-idf values, introduced in Equation 1). Geographic ranking, in contrast, is often approximated by geometric or geographic measurements, such as Euclidean distance, overlap or direction. A simple implementation of geographic ranking computes the relative overlap a spatial query and spatial footprints of documents (Larson and Frontiera 2004).

Evaluation. Mandl (2011), in a review on evaluation techniques of GIR Systems, distinguishes four relevant types of evaluation:

- Evaluation of the *component level*, which focuses on the implementation of particular components of the retrieval system, such as the indexing or ranking.
- On the *system level* the performance of the sum of all components is tested.
- *User-System-Interaction* applies an evaluation mostly at the level of the user interface, testing its suitability.
- On the *user performance level* the abilities and expertise of the user are incorporated in evaluation to see if it has significant impact.

We are mainly interested in evaluations on the system level, where the performance of the whole GIR system is evaluated against a baseline system. In traditional IR this is referred to as the *Cranfield* model for evaluation (as described by Borlund 2003).

Studies on search engine logs have shown that up to 18% of all queries contain spatial information (e.g. Gan *et al*. 2008), which suggests that using geographic intelligence to deal with the spatial dimension of a query should clearly improve information retrieval.

IR outperforms GIR. The most extensive evaluation initiative of a GIR system so far was GeoCLEF which ran from 2006 to 2008, with 33 research groups involved, 505 experiments submitted and over 100,000 human relevance judgments generated (Mandl *et al.* 2008). Throughout all GeoCLEF tracks it could not be shown that the incorporation of spatial indexes and rankings could outperform a simple text base line (Mandl 2011). The comparison of different GIR systems, and different queries in particular, is highly complex. Queries are multidimensional since they incorporate a variety of implicit contextual and spatial parameters, such as the publicity of a topic, the level of detail of the spatial element or the topological complexity of the spatial relations. Li et al. (2006) could for instance show that some query results benefit from the incorporation of spatial indexes and geographic intelligence. The reason this

could not be shown in GeoCLEF, they argue, is that GeoCLEF tends to use too simple queries, for instance containing spatial information on city or country level. On this spatial granularity level GIR could not outperform string based indexes.

GIR outperforms IR. Aside from GeoCLEF, there are examples of GIR evaluations where simple IR systems could be outperformed. Examples are SPIRIT (Purves *et al.* 2007) or Palacio et al. (2010). In SPIRIT the responses of a string and a spatial search were compared for 38 queries, incorporating different spatial relations and locations of different granularities. The precision values (i.e. relative number of correctly retrieved documents, see following paragraph for more information on evaluation measures) are based on relevance judgments of two annotators. The string search was clearly outperformed, with some 30 queries gaining higher precisions using GIR. The difference between textual and spatial search is most obvious for queries containing complex spatial relations, such as *near* or *within distance of*. Palacio et al. (2010) found that a GIR that incorporates all three dimensions of geographic information[15], namely textual, temporal and spatial dimension, could improve the state of the art retrieval system by some 75%. They could also show that "the three dimensions are not redundant, but they complement each other" (p. 105).

Precision and Recall. Retrieval is often evaluated using *precision* and *recall* values, where precision is the relative number of correctly retrieved documents (i.e. true positives), often calculated for the top X results - e.g. *p@10: 80%* means that 8 out of the 10 top ranked documents are relevant. Recall, on the other hand, is the relative number of correctly retrieved documents compared to the number of all correct documents available. The range of precision values for spatial queries, retrieved in former GIR initiatives is broad. In SPIRIT (Purves *et al.* 2007) they tested the retrieval performance on the basis of 38 queries and gained precisions as summarized in Figure 8.

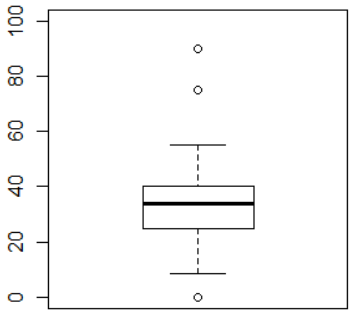


Figure 8. Precisions for 38 spatial queries summarized from SPIRIT (Purves *et al.* 2007, pp. 736–737)

[15] The notion that geographic information consists of three dimensions is used by the authors. We do not necessarily agree with this definition.

Recall values are only rarely given in GIR literature. The calculation of recall requires knowing the relevance of each document in the corpus for each query. This is often only feasible if the corpus is associated with metadata, or fairly small. Having both precision and recall, they can be combined to an overall accuracy measure. A widespread measure in IR is the F_1 value:

Equation 2. F_1 as a means for computing accuracy from recall and precision values.

$$F_1 = 2 \times \frac{recall \times precision}{recall + precision}$$

Summary Geographic Information Retrieval:

- GIR is a combination of methodologies from GIS and IR.
- Geographic information is particular in an information retrieval context because of toponym ambiguity, the influence of spatial granularity and the topological complexity of most spatial relations.
- Spatial indexing allows effective retrieval of spatial information.
- Spatial ranking allows ranking of spatially retrieved documents using geographic criteria (i.e. geographic relevance).
- IR usually outperforms GIR if queries contain simple geographic information (e.g. GeoCLEF).
- GIR usually outperforms IR if queries contain complex spatial relations or different types of information, such as topical, spatial and temporal specifications.
- Performance of a retrieval system is usually evaluated using *precision* and *recall* values. Recall can only be calculated if all documents are tested for relevance for each query. This information is most often not available.

2.2.2 Ambiguity and Toponym Disambiguation

"Georeferencing by placename (aka feature name) is the most common form of referencing a geographic location […]" (Hill 2009, p. 91). Thus, the linking of text documents to space is usually processed by grounding toponyms. Leidner and Lieberman (2011) sketched a workflow of the steps required for grounding toponyms from text (Figure 9).

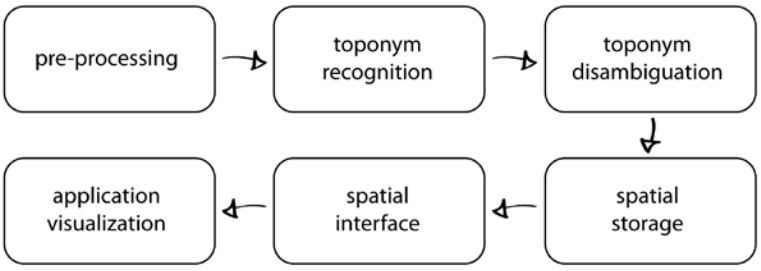


Figure 9. Model for grounding toponyms from text (modified from Leidner and Lieberman 2011)

In the following we will mainly focus on the two steps *recognition* and *disambiguation* of toponyms.

Toponym Recognition. The recognition of toponyms from text, also referred to as *toponym lookup* (Clough 2005), is often implemented as a token by token comparison between the text document and entries from a *gazetteer*, which is a list of toponyms with associated coordinates and a limited amount of additional information (Hill 2009). The output of toponym recognition consists of a set of tokens that have similar spelling to known toponyms (i.e. *potential toponyms*). There are many approaches that use gazetteers in toponym recognition including Purves et al. (2007), Amitay et al. (2004), Li et al. (2003) and Smith and Crane (2001). The result of a gazetteer based approaches clearly depends on the level of detail of the gazetteer, such that with increasing level of detail - a high level of detail is for instance needed to resolve fine spatial granularity information - the number of false positively recognized toponyms increases (an effect that is described in the next paragraph). More recent approaches aim at minimizing the influence of gazetteers, and gazetteer size in particular, by for instance combining gazetteers with machine learning algorithms for recognizing toponyms (e.g. Martins *et al.* 2010).

Toponym Disambiguation. Toponyms are often ambiguous such that the result from toponym recognition has to be disambiguated. Figure 10 gives an example of one type of toponym ambiguity. A sentence containing the toponym *New York* could be referenced to at least 10 different populated places, located in North America, Africa and Europe.

Figure 10. Populated reference locations to the toponym *New York* from Geonames.

Types of Toponym Ambiguity. Leidner (2007) lists three types of toponym ambiguity, namely *discord*, *non-specificity* and *linguistic ambiguity*. Discord ambiguity occurs if different groups of people or agencies disagree on the location of a particular toponym. Non-specificity ambiguity can occur if the delivered information is not sufficiently specified, such that one cannot resolve only one possible location

(e.g. north of London). Linguistic ambiguity, which is most important in this thesis, is grouped into three sub-types: *morpho-syntactic*, *feature type* and *referential ambiguity*:

- Morpho-syntactic ambiguity: a token constitutes a toponym and, in parallel, a non-specific, non-named entity concept. Examples are *bath*, that could be a place north of London or the object *bath*. This type of ambiguity is particularly difficult to resolve in languages with capitalized nouns, such as German.
- Feature type ambiguity: the same token refers to different geographic feature types (*Zürich* the canton and *Zürich* the city).
- Referential ambiguity: One toponym can be referenced to several locations (e.g. 25 mountains have the name *Schwarzhorn* in Switzerland, c.f. Figure 10).

For the sake of simplicity we will use the terminology introduced by Amitay et al. (2004) and refer to ambiguity as either *geo/geo*, or *geo/non-geo ambiguity*. Geo/non-geo ambiguity is equal to *morpho-syntactic ambiguity*, whereas geo/geo ambiguity covers both *feature type* and *referential ambiguity* (e.g. Figure 10). Investigations of corpus data have shown that 67% of toponyms in an average text are geo/geo ambiguous (Garbin and Mani 2005), and 17% of all toponyms in newspaper articles are geo/non-geo ambiguous (Leveling and Veiel 2007).

Toponym ambiguity affords *toponym disambiguation* - or *toponym resolution*, as Leidner (2011) calls it. Ambiguities that occur due to sparse information (non-specificity) and human disagreement (discord) are, to our knowledge, not covered in automatic approaches of toponym disambiguation.

Disambiguation Approaches. Buscaldi (2011), in a review on *toponym disambiguation*, distinguishes three approaches, namely *map-based*, *knowledge-based* and *data-driven*. Map-based approaches assume *geometric-minimality* (Leidner 2004), where the spatial extent of a document footprint is minimized. Thus, geometric-minimality reflects Tobler's first law of geography, where proximity is considered a proxy for similarity (Tobler 1970). However, geometric-minimality is very sensitive to *outlier* locations, for instance caused by sudden changes of the subject of a description. Imagine for instance a detailed description of a particular ascent of a mountain. The geometric-minimality heuristic might be valid as long as the writer describes the *ascent*, and thus lists toponyms along the trail. However, as soon as the writer changes the subject, for instance by recalling a past ascent, he will no longer obey the geometric-minimality heuristic and suddenly change the spatial context. Map-based approaches are often used. Examples are Smith and Crane (2001) and Buscaldi and Magnini (2010). Smith and Crane (2001) limit the extent of document footprints by ignoring outlier locations, defined by a threshold distance. Buscaldi and Magnini (2010) use metadata on the 'real' origin of a description in order to decide if a toponym

location is considered in the footprint. Such metadata, however, is only rarely available (e.g. Wikipedia). Often, the geometric-minimality assumption is used in combination with additional heuristics, such as knowledge-based approaches.

Knowledge-based approaches apply toponym information, which is for instance available from gazetteers. Population count is frequently applied, assuming that higher population counts increase the probability that a particular toponym location is meant in text (e.g. New York, as represented in Figure 10, will always be resolved as the one New York in the State *New York*) (e.g. Amitay *et al.* 2004, Overell and Rüger 2008). Buscaldi and Rosso (2008) resolve geo/geo ambiguity by using information from the WordNet[16] ontology, namely *Synsets*, i.e. lists of synonyms (e.g. London, Greater London, British Capital), and *semantic relationships*, such as *meronymy* (part-of) or *hypernymy* (is-a). The WordNet information is used to perform disambiguation by computing *conceptual density* (e.g. Agirre and Rigau 1996), which is the correlation between the sense of a word, gathered from WordNet ontology, and the context in which the word occurs in text, gathered from neighboring terms. Bensalem and Kholladi (2010) perform geo/geo disambiguation using a minimality heuristic compound of geometric and *semantic minimality*. Semantic minimality is calculated from *arborescent proximity*, which is the hierarchical distance between toponyms in the *tree of world places.* The tree of world places is a hierarchical structure of locations mostly using administrative classification, such as continent, country or state. An interesting source of knowledge, rather than a knowledge-based disambiguation approach, is described in Alazzawi et al. (2012). They describe an approach for retrieving place relevant information for locations stored in a gazetteer from *DBpedia*[17]. This information can then be used to answer questions such as *What can I do there?* and is thus of potentially value in a disambiguation context.

Strictly speaking, data-driven approaches are a sub-set of knowledge-based approaches, with the particularity that toponym knowledge is used in machine learning. Martins *et al.* (2010) describe an approach for performing toponym disambiguation using a *Hidden Markov Model* to annotate place references and *Support Vector Regression* in order to perform disambiguation. The feature space of toponyms, used to train and test the machine learning algorithm, is populated by six measurements, namely *Levenshtein distance* between known toponyms and tokens in the text, population counts, number of alternative names of toponyms, spatial distance, size of convex hull, and size of concave hull. Data-driven approaches were only recently applied in toponym disambiguation and usually suffer from a lack of tagged data (i.e. gold standard). Additionally, they only poorly classify unseen toponyms (Buscaldi 2011).

[16] wordnet.princeton.edu
[17] dbpedia.org

Often, map- and knowledge-based approaches are used in combination. The *Web-a-Where* GIR System, introduced by Amitay et al. (2004), does, for instance, combine the map-based *geometric minimality* assumption with the knowledge-based *largest population* heuristic. The disambiguation approach introduced by Martins *et al.* (2010), which is described above, incorporates all three approaches, map-based, knowledge-based and data-driven.

Summary Toponym Ambiguity and Disambiguation:

- Linking text to spatial footprints is called *geoparsing*.
- Geoparsing consists of *toponym recognition* and *toponym disambiguation*.
- Toponym recognition is often performed through *toponym lookup*, i.e. the comparison of entities in a gazetteer with words occurring in text. Thus, the level of detail of the gazetteer has crucial impact on the lookup output.
- Toponym disambiguation is motivated by toponym ambiguity which in this thesis is divided into geo/geo (20 instances of the mountain *Schwarzhorn* in *Switzerland*) and geo/non-geo ambiguity (*Berg* can be a toponym as well as a generic noun, i.e. mountain).
- Toponym disambiguation is performed using *map-based* or *knowledge-based* approaches. *Data-driven* approaches are a third category that use map- or knowledge-based information and apply machine learning.
- Map-based approaches often assume that the footprint of a document has to be of minimum extent (i.e. *geometric-minimality*).

2.2.3 Disambiguation of Natural Features

Leidner (2007) argued that toponym disambiguation has often only focused on populated places, typically of coarse spatial granularity level. Brunner and Purves (2008) conducted an investigation in Switzerland on the relationship between geographic feature types of toponyms and occurrences of referent ambiguity with the result that only some 5% of populated places are ambiguous, whereas more than 40% of all toponyms in the gazetteer are geo/geo ambiguous (Figure 11).

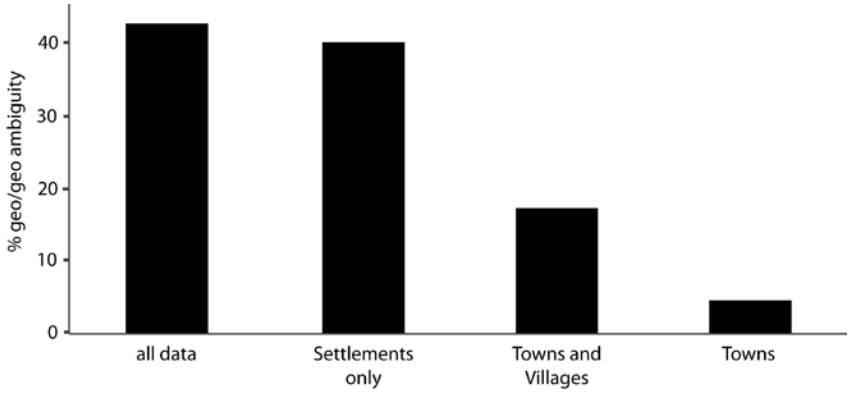


Figure 11. Referent ambiguity for toponyms of different feature types in Switzerland (Brunner and Purves 2008).

Thus, most approaches to toponym disambiguation have concentrated on a rather simple set of toponyms, from an ambiguity point of view. In contrast, a corpus consisting of natural landscape descriptions that

contains references to toponyms of natural feature types is thus assumed to be more prone to geo/geo ambiguity. As well as above-average geo/geo ambiguity, the disambiguation of toponyms of natural feature types, such as mountain, hill or hamlet, is complicated by sparse toponym information, such that most approaches described in Buscaldi (2011) are not applicable (e.g. largest population). Sparse toponym information is not an inherent property of toponyms referring to natural features, however, most natural features are of rather fine spatial granularity and often located far from densely populated places, and thus, not very well known. As a consequence they cannot be disambiguated using state of the art disambiguation approaches. Consider for instance the above mentioned approach of Alazzawi et al. (2012) of gathering place related information from *DBpedia*. This approach will certainly fail in gathering relevant information for all 350 instances of toponyms called *Rüti* in Switzerland (Figure 12, left).

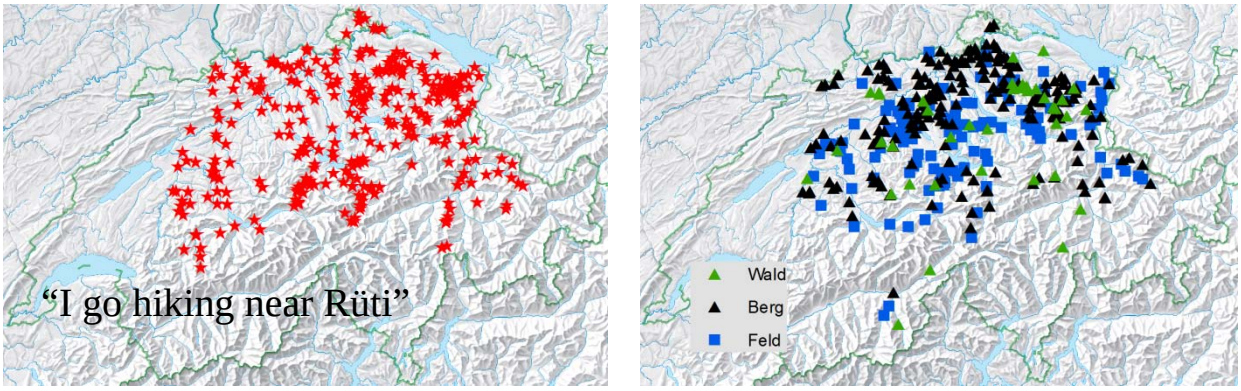


Figure 12. Geo/geo and geo/non-geo ambiguity of fine spatial granularity. Demonstrated with referent locations of Rüti (left, geo/geo) and referent locations of the three terms/toponyms *Wald* (forest), *Berg* (mountain) and *Feld* (field) (right, geo/non-geo and geo/geo).

The only disambiguation heuristic that is independent from effects on fine spatial granularity is geometric-minimality, as introduced above. We already mentioned that the application of geometric-minimality suffers from the extreme flexibility of language. As an additional constraint for applying geometric-minimality, Brunner and Purves (2008) could show that referent locations of geo/geo ambiguous toponyms in Switzerland are spatially autocorrelated, such that their average distance is less than half of the distance expected between random samples (45 vs. 100km). This result is independent of language region and has further implications on the use of geometric-minimality as a single disambiguation heuristic. Some examples of spatial autocorrelation can be observed in Figure 12 (left), where often several *Rütis* are located in the same valley. Depending on the spatial extent of an individual document, it might thus not be possible to unambiguously resolve one referent location.

We observed that disambiguation approaches often focus on geo/geo ambiguity, whereas geo/non-geo ambiguity is usually not discussed as a major problem. In Figure 12 (right) we show referent locations of

three toponyms that are geo/non-geo (and geo/geo) ambiguous. These three terms often occur in natural landscape descriptions, which makes disambiguation critical and difficult.

From this review of approaches to resolve toponym ambiguity, and from the focus on natural features in particular, it is clear that natural landscape descriptions represent a new challenge to toponym disambiguation, both, in terms of over-average referent ambiguity, and because most known approaches are not applicable.

Summary Disambiguation of Natural Features:

- Toponym disambiguation has so far mainly concentrated on documents referring to toponyms of coarse spatial granularity, such as cities or country names.
- Toponyms of finer granularity level, for instance referring to natural feature types, such as mountains, hills or hamlets, are prone to pronounced toponym ambiguity.
- Additionally, ambiguous referent locations to a toponym are significantly autocorrelated (for Swiss and British toponyms).
- Thus, disambiguation of natural landscape descriptions, containing references to natural feature types, is complex, since we often lack explicit toponym knowledge and since we are usually confronted with over average toponym ambiguity (geo/geo).
- This suggests the use of map-based approaches, which, however, is of limited applicability since ambiguous toponyms are autocorrelated.
- For the above mentioned reasons, disambiguation of natural landscape descriptions must be considered a new challenge in geoparsing.

2.2.4 Digital Humanities and Literary GIS

Digital Humanities. A very recent field of science that primarily aims at extracting information from digitized text is the *digital humanities*, where digital input data is used to answer research questions from human and social sciences (e.g. Berry 2012). Nature, in an editorial article on an approach that uses millions of e-mails as input to scientific investigations states that "[s]uch research could provide much-needed insight into some of the most pressing issues of our day, from the functioning of religious fundamentalism to the way behavior influences epidemics…" (Nature 2007, p. 637). Another example is published in Michel et al. (2011), where the authors aim to resolve temporal plots of cultural trends, covering the last two centuries. The information is retrieved from some 5 million digitized books, which is about a fourth of all Google books. Interestingly, geography and geographic information in particular is ignored in most prominent investigations associated with digital humanities. An exception is literary GIS, a domain that is often associated with the digital humanities.

Literary GIS. Moretti (1998), in his seminal book *Atlas of the European Novel 1800-1900* considers maps as an analytical tool in order to "bring to light relations that would otherwise remain hidden" (p.3). Moretti (1998) is often regarded as an early example of literary GIS or, *literary geography*, as he named

it. Cooper and Gregory (2011) report on a recent literary GIS approach. They mapped two novels by applying manual annotation. The strength of Cooper and Gregory's "Mapping of the English Lake District […]" is that mapping is only a preprocessing step in order to allow follow up investigations on the semantic content of the two novels. They call the mapping product a *macro-map*[18]. Cooper and Gregory (2011) argue that "mapping in literary studies, has frequently become synonymous with a way of reading rather than cartography." (p.91). Different visual variables, such as symbology or color schemas, are used for transmitting different information, for instance for distinguishing visits from mentioned places or for representing the mood of the author when describing a particular landscape.

Piatti (2008) in her book *Die Geographie der Literatur* is not primarily interested in spatial representations of text, but in the different relationships between space and literature. The relation can be *realistic*, such that for instance an existing city is described in great detail, or *fictional*, such that fictional stories are associated with existing places or that new places are invented for the sake of the story. Piatti's (2008) investigation is motivated by a fundamental research question, namely on how literature uses space. The georeferencing is manually performed and requires detailed geographic and literary knowledge. Consequently, Piatti's approach of drawing maps from text is very time consuming.

Approaches associated with literary GIS, usually perform manual annotation in order to link text with space and thus only consider a limited number of documents. This is clearly not state of the art in GIR, as discussed above, where automatic geoparsing allows the processing of thousands of documents. The reason for still discussing approaches from literary GIS, is the use the mapped text documents for further analysis. Mapping is considered a preprocessing step in order to conduct more in-depth content analysis. We have never seen this in GIR approaches, where mapping is usually a means for designing applications that allow the retrieval of locally relevant information.

Summary Literary GIS:

- Digital humanities is a fast growing field where digital human sourced data is analyzed for answering research questions from human or social sciences.
- Geography does not play a major role in digital humanities.
- Literary GIS is often mentioned as one representative of geography in the digital humanities.
- Literary GIS creates maps from text, mostly through manual annotation, which has consequences on the number of documents that can be processed.
- Compared to GIR, literary GIS uses maps as a building block for further investigations, for instance on the semantic content of descriptions, or the fundamental role of space in literature.

[18] The macro-map is a cartographic representation of all locations listed in the two novels that are investigated.

2.2.5 Critical GIS

In the introduction of this chapter on geographic information retrieval we argued that the data avalanche, in terms of increasing volume of digitally available text, has consequences on social and human sciences, in such that they are urged to increasingly rely on digital information and processing for answering research questions.

One potential role of GIS, discussed in the previous chapter, is the mapping of text. This is considered an important precondition for further geographic interpretations. However, the question emerges about whether GIS is ready to capture, store and process human sourced information. In the discussion on landscapes and landscape concepts we emphasized that landscape features, such as mountains, have undetermined boundaries and that different definitions for mountain coexist in individuals, groups or cultures (§2.1.5). We linked this uncertainty to the debate on vagueness, and concluded that vagueness is an unsolved challenge.

Additionally, Gary Lock (2010) concluded that the use of digital spatial technologies in the humanities introduces a new tension, which could be added to John Wylie's (2009) understanding of landscape. Wylie used a set of four tensions to describe the complex role of landscape: *distance and proximity*, *observing and inhabiting*, *eye and land* and *culture and nature*. Lock argues that digitizing comes at the cost of simplification of the real world, and that thus, *interpretation* must be considered a fifth tension. The role of interpretation becomes more central, since statistical outputs (e.g. regressions, correlations or dependencies), gathered from large digital input data are less self-explanatory, and thus vague and ambiguous, compared to conclusions drawn from individual observations in the field or gained from classroom experiments.

A critical view on using GIS methodologies in order to represent human sourced information is reflected in the *critical GIS* literature. The foundation of critical GIS is Pickle's collection of essays *Ground Truth* (Pickles 1994). Pickle argues that GIS is a corporate product to solve corporate problems, such as logistics or market analysis. Thus, GIS is rested on a positivist epistemology and employs a linear logic to the world, implying for instance sharp borders, formal definitions or exact numeric values, which do not adequately represent complex, real world problems central to social and human sciences. Pickles argues that GIS is designed to answer *what* and *where* questions, instead of approaching *why* causalities.

A positivist would argue that the *why* could be deduced from *what* and *where* information. A critical humanist, however, would insist that this knowledge is not universal and thus clearly depends on the perspective of the observer, for instance reflected by research in ethnophysiography (§2.1.5). Bodenhamer

et al. (2010) summarized several critical facets by stating that “GIS privileged a certain way of knowing the world, one that valued authority, definition and certainty over complexity, ambiguity, multiplicity, and contingency, the very things that engaged humanists” (Bodenhamer *et al.* 2010, p. ix). This criticism highlights the limits of the applicability of a GIS to social and human sciences, but does not disallow the use of GIS as a tool.

Recent applications of GIS in the humanities do not aim at holistically representing human information. This is for instance reflected in the work discussed under the umbrella of literary GIS (§2.2.4). GIS is rather used as an additional source of knowledge. Bodenhamer et al. (2010) stated that “[w]e are drawn to issues of meaning, and space is a way to understand fundamentally how we order our world” (p.14). One example is reported in Fairclough (2006), where a GIS is used as an archeological tool for resource management, i.e. *Historic Landscape Characeterization*[19] (HLC). HLC has been argued to be "a way of going beyond intuition to get beneath the skin of a place and look at its essential qualities and character"[20]. HLC was originally motivated by the need for archiving historic landscape maps in Britain (e.g. Herring 2009) and grew to be a tool for representing the *historic character* of a landscape.

Summary Critical GIS:

- Human sourced data is often ambiguous and vague.
- Using digital information for answering social or human research questions can change the role of *interpretation*.
- GIS often follows a positivist paradigm by assuming sharp boundaries and clear-cut definitions.
- Critical GIS serves as a foundation for clarifying the role of GIS and potential common grounds with social and human sciences.

2.2.6 Summary

In this chapter we had a closer look at approaches to linking text to spatial footprints and maps. Such approaches are covered by the two research topics *Geographic Information Retrieval* (GIR) and *literary GIS*. Both fields use different methodological frameworks and have different goals, which will be quickly summarized below.

Text as Map. Geographic information is often **not explicit** in human sourced data, such as text. An example is the use of toponyms in written natural language, which are seamlessly embedded in text. Such implicit information must be extracted, before it can be used for conducting geographic investigations. **Literary GIS** describes approaches where geographic information, contained in landscape descriptions, is

[19] www.english-heritage.org.uk/professional/research/landscapes-and-areas/characterisation/historic-landscape-character
[20] www.archaeologyuk.org/conservation/planninguide

manually annotated and represented in maps, i.e. **macro-maps**. Literary GIS considers macro-maps as an important means for deducing new, unseen information from descriptions. Thus, macro-maps are used for conducting semantic content analysis of text descriptions. However, drawing interpretations from results that are deduced from human sourced data, as is state of the art in the **digital humanities** and literary GIS is critical and care must be taken. Such issues are covered in **critical GIS**, for instance by exploring and discussing the limitations of a GIS.

Manual annotation introduces a clear limitation to approaches in literary GIS in terms of the number of documents under consideration, such that a large corpus, consisting of some thousand documents, could not be processed. A more information science driven approach for linking text to spatial footprints is described in the **GIR** literature, where **geoparsing** allows the automatic grounding of toponyms from text. However, the main focus of GIR is not on spatial representations and interpretations of text, but spatial indexing and ranking. Indexing and ranking is performed for facilitating the resolution of spatial queries in information retrieval (i.e. GIR). This is clearly reflected by the type of text documents that is usually considered in GIR, which reflects the major everyday human information need.

Geoparsing of Natural Landscape Descriptions. Geoparsing has mainly focused on text documents that describe space in **coarse spatial granularity** or methodologies that only support the retrieval of coarse spatial information. This introduces a bias towards descriptions that refer to populated or well-known places, such as cities or countries, and an under-representation of descriptions of unpopulated and natural landscapes. Geoparsing is distinguished into toponym recognition, often in the form of **toponym lookup**, and **toponym disambiguation**. The result of toponym lookup is clearly dependent on the level of detail and spatial coverage of the **gazetteer** which is used. The resolution of fine spatial granularities from text, such as those present in natural landscape descriptions, affords the use of a detailed gazetteer. This automatically introduces more false positives. Toponym disambiguation, on the other hand, aims to resolve unambiguous toponym locations from all toponyms that were recognized in the lookup. Toponym ambiguity either has the form of **geo/geo** (i.e. one toponym has several referent locations, e.g. some 25 *Schwarzhorn* in *Switzerland*) or **geo/non-geo ambiguity** (i.e. the wording of the toponym is also used for common nouns, e.g. *Berg* occurs as a toponym and also refers to the generic term mountain). State of the art approaches to toponym disambiguation are either knowledge- or map-based. **Knowledge-based** approaches use information that is explicitly available on a toponym level, such as population counts or administrative function. However, such information is usually not available for toponyms referring to natural features. **Map-based** approaches usually make the assumption that the extent of a document footprint is to be minimized. This assumption does not necessarily match with the intention of an author and is of limited applicability due to autocorrelation of **geo/geo ambiguous** toponyms. Thus, natural

landscape descriptions constitute an unsolved challenge to geoparsing and afford the introduction of new geoparsing heuristics.

In the following chapter we will use the discussed body of literature, related to landscape research and the linking of text to spatial footprints, in order to resolve four research gaps. These research gaps are described and will then be used to motivate the investigations, which built the core of this thesis.

2.3 Research Gaps and Questions

We resolved four research gaps from the previous two chapters on the state of the art in landscape research and the linking of text to space. We are aware that these gaps represent a subjective selection, and that more or different research gaps could be resolved from the same scientific context.

RG I: The automatic linking of natural landscape descriptions to space. State of the art approaches in GIR, aiming at linking text to spatial footprints, usually focus on descriptions containing references to populated places of coarse spatial granularity. *In order to link natural landscape descriptions to spatial footprints, we must therefore introduce a new approach, which is optimized on toponyms referring to natural features of often fine spatial granularity.*

RG II: Automatic "macro-mapping" of a whole corpus of natural language documents. Macro-mapping is a term introduced in literary GIS. There, the spatial representation of descriptions is used as a means to deduce information, additional to insights gained from close reading. Literary GIS usually incorporates individual novels, rather than whole corpora, and, importantly, geographic information is manually annotated. *Macro-mapping would therefore benefit from automatic methods of extracting geographic information.*

RG III: Investigation of landscape concepts from descriptions composed in natural language. Sample sizes in Ethnophysiography and empirical investigations on landscape features are usually small and of limited spatial and temporal coverage. *Approaches merging research questions from Ethnophysiography with methodologies capable of automatically gathering information from landscape descriptions from (historic) books, user generated contents or gazetteers, could therefore help to significantly increase sample sizes, and extend spatial and temporal coverage of landscape studies.*

RG IV: Using folksonomy to capture local subtleties in landscape concepts. There are many approaches that suggest the use of formal ontologies to store and structure geographic information. We

argued that such approaches are only of limited applicability since vagueness is inherent to many types of geographic information. This is particularly true for landscape features. Additionally, we criticize using top down approaches, such as a group of experts defining a finite set of terms which is then considered the valid taxonomy. This often poorly matches people's concepts of their environment indicated, for instance, by the fact that these terms would only rarely occur in natural language. *We will therefore rely on a bottom up approach, called folksonomy - in our case a spatial folksonomy for resolving local subtleties in landscape concepts. It could be considered a separate research gap that we deduce a folksonomy from natural language texts rather than from lists of tags in social media.*

These four gaps are reflected in the three research questions, introduced at the beginning of this thesis:

> RQ 1: *How can natural landscape descriptions be linked to space, with particular consideration of ambiguity in toponyms referring to natural features?*
>
> RQ 2: *How can local landscape concepts be captured from descriptions, under consideration of the vagueness associated with geographic concepts?*
>
> RQ 3: *Does the introduction of methods aiming to incorporate vagueness and ambiguity result in improvements in retrieval effectiveness for geographic information retrieval?*

2.4 Methodological Approach

The remainder of this thesis is structured into two consecutive topics as sketched in Figure 13. In the introduction we argued that the availability of large compilations of digitized landscape descriptions is important for geography for mainly two reasons. Firstly, as represented by topic 1, geography is important for gathering a first overview of the data. We called this the role *of* geography. Secondly, we argued that the information in this data might be crucial for contributing to fundamental geographic research questions. We called this the role *for* geography, and cover it in topic 2.

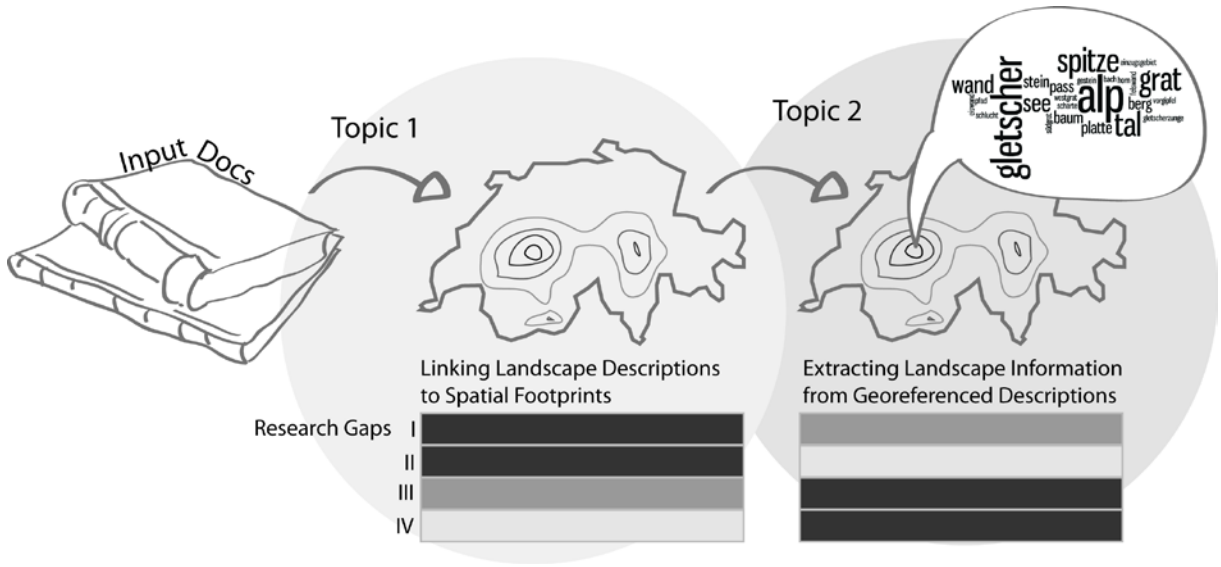


Figure 13. Structure of the thesis, broken down into two topics, associated with research gaps - dark grey colors emphasize strong association.

In the following we quickly summarize the content of the two topics, namely the *linking of landscape descriptions to spatial footprints*, and the *extraction of landscape information from georeferenced descriptions*, and put them into the context of the research gaps described in §2.3.

2.4.1 Topic 1: Linking Landscape Descriptions to Spatial Footprints

The first topic explores linking textual descriptions of natural landscapes to spatial footprints, which is, simply put, transforming text into a map (Figure 13, Topic 1). Literary GIS suggests the term macro-mapping for representing text as a map, which they associate with introducing a new layer of information and interpretation, compared to traditional close reading. The automatic transformation of a whole corpus of descriptions into a large set of spatial footprints is a significant contribution to the field of literary GIS (RG II).

Linking natural landscape descriptions to spatial footprints is a particular challenging task in GIR, and therefore resolved as an individual research gap (RG I). The resolution of toponym ambiguity usually incorporates toponym information, such as population counts, which is often not available for toponyms referring to natural features, such as mountains, rivers, hills or single hamlets (§2.2.2). For this reason we aim at finding other sources of toponym information, and thus, motivated by literature on the nature of geographic features (§2.1.8), replace missing toponym information through topographic characteristics, such as slope, curvature, relative drop or texture.

Output. The output of topic 1 consists of a detailed description and evaluation of the new approach for linking landscape descriptions to spatial footprints. Additionally, we produce spatial footprints for some ten thousand documents, which can be represented as a macro-map. Finally, we will use the spatial footprints for computing a spatial index that is central for the investigations associated with topic 2.

2.4.2 Topic 2: Extracting Landscape Information from Georeferenced Descriptions

At an early stage of this thesis we decided to consider digitized landscape descriptions, as input data for conducting investigations on landscape concepts. Using textual landscape descriptions, rather than conducting user experiments or field walks, is a significant contribution to the state of the art in landscape research, in terms of spatial and temporal coverage (RG III). We are aware that large coverage comes at the cost of level of detail, which is clearly higher in ethnographic studies.

Working with natural language descriptions, instead of human subjects, requires georeferencing (Topic 1). In Topic 2, we go one step further and use the georeferenced descriptions in order to extract information on the description of individual natural landscapes. Landscape information is approximated from representation of natural features in text, where natural features are resolved through manual annotation.

Output. The output of topic 2 is a spatial folksonomy – i.e. a georeferenced and weighted vocabulary of natural features gathered from text. The spatial folksonomy is used for different purposes. The comparison of the spatial folksonomy with existing land cover classifications gives us the means of discussing the advantages and disadvantages of a bottom up data structure (i.e. folksonomy), compared to rather formal top down taxonomies. This is a contribution to RG IV.

Chapter 3 Data Description

In this section we will introduce the data sets that are used in the following investigations. We separately discuss *gazetteer* and *corpus data*, *digital elevation models* and *landscape classifications*. The description of the data in a separate chapter should emphasize the importance of the input data for this thesis, in particular the two sources of landscape descriptions. The characteristics of the data can have major impact on the outcome of all investigations.

3.1 Gazetteer Data

As a gazetteer we mainly use *Swissnames*, a gazetteer of all toponyms found on *Swisstopo*[21] maps at scales of 1:25,000-1:500,000, with a total of more than 156,000 entries. Since the original motivation to compile Swissnames was cartography, and placing labels onto topographic maps in particular, toponyms are referenced to the geographic point coordinates where the particular toponym is found on the map (Figure 14).

Figure 14. Example of a Swiss topographic map of the scale 1:25,000. The red stars are labeled Swissnames referent locations for the respective toponyms in the map.

[21] Swisstopo is the Swiss federal mapping agency, www.swissnames.ch

This has an effect on precision. Features of small spatial extents, or with well-defined center points, usually have precise locations (Figure 14, e.g. *Kleines Fiescherhorn* or *Fiescherjoch*), whereas areal or linear features of rather big extents often fall somewhat short in terms of spatial representation (e.g. *Fieschergrat*). We are aware of this limitation. However, Swissnames is the most extensive gazetteer available for Switzerland.

Toponyms in Swissnames are associated with a set of descriptive attributes. Most interesting for our work is a feature type classification where each toponym is classified into one of 61 available feature types. The finite list of feature types is defined by Swisstopo. Often, the feature types do not correspond with terms used in everyday communication. There is for instance no *Berg* (mountain) in Swissnames. Instead, we identify labels such as *HGifpel*, *KGipfel*, etc.

According to the feature type classification, more than 50% of all toponyms in Swissnames refer to *natural* features. Figure 15 visualizes the (logarithmic) frequency distribution of feature types as tag clouds. We thus classified feature types as either being natural or artificial. There are slightly more natural features in Swissnames. However, artificial features are represented by a larger number of different feature types. The two most prominent feature types are *Flurname* and *Einzelhaus*. *Flurname*[22] has no equivalent in English. In German they are used to refer to small spatial extents such as fields or moors in natural landscapes. *Einzelhaus* is used for the class of toponyms labeling single buildings that are salient in landscape. *Flurname* and *Einzelhaus* both refer to features of small spatial extents.

natural features (n = 83507) artificial features (n = 73248)

Figure 15. Tag clouds from logarithmic frequencies of natural (left) and artificial (right) feature types in Swissnames. (Source: Swissnames, www.wordle.net)

[22] We use the term Flurname as a class of toponyms, indicating that these toponyms refer to small landscape features. Unlike other classes of toponyms that frequently refer to particular feature types, such as mountains or valleys, Flurnamen cannot be considered a feature type. Flurnamen can be of very different feature type, ranging from fields and meadows to lakes, hills and mountains. The shared characteristic is that these features mostly are of small spatial extent.

Table 2 contains information on feature types that are discussed in some of the following investigations.

Table 2. Swissnames feature types, discussed in some of the following investigations.

type	translation	count
Flurname	related to field name	54980
Bach	stream	3960
Fluss	river	399
KBach	small stream	1004
GGipfel	prominent mountain	866
HGipfel	major mountain in a region	165
Grat	mountain ridge	1440
Huegel	hill	2543
Gletscher	glacier	730
GSee	small lake	53
KSee	lake	817
Wasserfall	waterfall	52
Quelle	spring	69
Weiher	pond	101
Sumpf	moor/marsh	191
GOrtschaft	big town	112

43% of all individual toponyms in Swissnames are referent ambiguous (i.e. more than one instance of the same name occurred within Swissnames). Referent ambiguity is not equally distributed over all feature types. Populated places seem to be less prone to referent ambiguity, compared to all other feature types (Figure 11, Brunner and Purves 2008). Only some 3% of all unique toponyms in Swissnames are geo/non-geo ambiguous, such that they are tagged as nouns in the TIGER corpus (§3.2.3).

3.2 Corpus Data

We used different corpus data. A corpus, in linguistics, is considered an often large set of annotated text documents (e.g. Marcus *et al.* 1993). A prototype example, central to many of our investigations and described below, is Text+Berg (Volk *et al.* 2010). In some cases we slightly broaden the linguistic concept of a corpus, for instance by calling a large set of tagged Flickr[23] images a corpus.

[23] www.flickr.com

3.2.1 Text+Berg

Text+Berg is a digitized collection of *Swiss Alpine Club*[24] (*SAC*) yearbooks dating back to 1864. In the version of the corpus we work with, a total of 134 yearbooks were present, each with around 80 articles and 300-600 pages (Volk *et al.* 2010). This is an equivalent of some 36 million tokens. Text+Berg has a broad topical focus, containing descriptions of classical and modern mountaineering, contemporary descriptions of many of the first ascents in the Swiss Alps, regular reports on the condition of Swiss glaciers and much more. The corpus is multilingual with articles mainly in German, but also in French and Italian. Before 1957 a majority of the articles in the yearbooks were written in German. Approximately 10% of all articles were written in French and only few in Italian. Since 1957 parallel yearbooks have been published in French and German. There is no obvious pattern evolving from early articles of French titles predominantly focusing on the French speaking part of Switzerland, neither do German articles only describe the Swiss German Alps.

Figure 16 shows an example of an article from 1900 with the title "*Bergfahrten im Clubgebiet*", describing different ascents that took place that year.


Das war eine Scene hier oben, ärger als in Dantes Hölle: Schneetreiben, Nebel, sausender, heulender Wind und Kälte. Da war's gestern

Darray, Grande Luis, Tour noir, Aiguille de la Neuva vom Portalet aus.
Nach einer Photographie der Collection *Attinger*.

auf der Aiguille du Tour doch etwas schöner, war mein erster Gedanke. Unter diesen Umständen ließen wir uns von dem Berg nicht lange „fecken", wie der Oberländer sagt, nahmen in aller Eile, ohne abzusitzen, einen kurzen Imbiß und stiegen, nach 5 Minuten Halt, auf der Westseite des Berges wieder ab zum **Col des Ravines rousses.** Es war ein steiler Weg

Figure 16. Extract of an article from 1900, written by A. Walker ("Bergfahrten im Clubgebiet", p.19).

We received the corpus in a digital, preprocessed format, as separate syntax-parsed XML files for each yearbook (Figure 17). The preprocessing, as well as the data compilation and digitization was performed by the Institute of Computational Linguistics of the University of Zurich[25]. Central to our purpose, the

[24] www.sac-cas.ch
[25] www.cl.uzh.ch

parsed format identifies individual articles and carries out part-of-speech tagging and lemmatizing on individual tokens (Sennrich *et al.* 2009). Since these methods are standard in computer linguistics, we assume that errors induced by this preprocessing are not significant.

Über die Jungfrau mit Ski, 1940
Otto Gerecht

Wiederum hatten wir ein paar prachtvolle Maitage im Oberland verlebt, hatten Schreckhorn, Agassizhorn und Grosses Fiescherhorn bestiegen.
[...]
Zuerst verschneite Felsen, dann ein Sprung über den Bergschrund, und schon rutschen und zirkeln wir den Hochfirn hinab.

art	sent	id	pos	lemma	word
35	13	0	ADV	wiederum	Wiederum
35	13	1	VAFIN	haben	hatten
35	13	2	PPER	wir	wir
35	13	3	ART	ein	ein
35	13	4	PIAT	paar	paar
35	13	5	ADJA	prachtvoll	prachtvolle
35	13	6	NN	Maitag	Maitage
35	13	7	APPRART	im	im
35	13	8	NN	Oberland	Oberland
35	13	9	VVPP	verleben	verlebt
35	13	10	$,	,	,
35	13	11	VAFIN	haben	hatten
35	13	12	NE	Schreckhorn	Schreckhorn
35	13	13	$,	,	,
35	13	14	NE	Agassizhorn	Agassizhorn
35	13	15	KON	und	und
35	13	16	NE	Grosses	Grosses
35	13	17	NE	Fiescherhorn	Fiescherhorn
35	13	18	VVPP	besteigen	bestiegen
35	13	19	$.	.	.

Traversing the Jungfrau with skis, 1940
Otto Gerecht

Once again we had a few glorious days in May in the Oberland. We climbed on Schreckhorn, Agassizhorn and Gross Fiescherhorn.
[...]
First snow covered rocks, then a spring over the bergschrund and we slide and thread our way down through the firn.

Figure 17. Example sentences from an article from Text+Berg, consisting of the original German text, a part-of-speech tagged version and an English translation (Derungs and Purves 2013).

3.2.2 HIKR

HIKR[26] is a non-profit website where users can publish reports on their outdoor activities. The basic idea is that people have one platform that suits different purposes, such as archiving, networking and sharing information. To date there is a total of some 50,000 articles, published by some 10,000 registered users on HIKR. Articles on HIKR are relatively short in length. The average length is 300 words, with 66% of all articles having between 100 and 500 words. This is an equivalent of 1 to 2 pages of text. Only 3% of the articles consist of more than 1000 words.

The majority of articles are written in German, with French, Italian and English also being represented. The articles are distributed all over the world with a clear focus on the *Alps*. We will concentrate on German articles that refer to Switzerland, which is given for some 25,000 articles.

Figure 18 shows an example of a HIKR article, consisting of metadata (box) and the text description.

[26] www.hikr.org

Figure 18. Example of a HIKR article[27], consisting of metadata and the text description

An interesting feature of HIKR articles is the associated metadata. Among other information, it consists of a hierarchical taxonomy of the region, the date, the type of activity and the associated difficulty, and some selected waypoints. The metadata information is explicitly added by the author. We are particularly interested in the activity classification and the waypoints.

The activity classifications distinguishes between *hiking*, *skiing*, *ski touring*, *mountaineering*, *climbing*, *ice climbing*, *mountain biking* and *snow shoe hiking*. The difficulty of each of the activities is taken from official schemas, the climbing and mountaineering difficulties are for instance gathered from an existing classification of the *Swiss Alpine Club*. Each article can be associated with several activities. The metadata information on the activity can be considered ground truth information on the topic of the description. If an article is tagged as a *mountaineering* article, the content of the description is assumed to describe mountaineering, however, without the necessity of explicitly mentioning *mountaineering* in text.

The waypoints can be considered ground truth information on the spatial footprint of the description. The most important toponyms are explicitly listed and associated with geographic coordinates. Again, the waypoints can be considered ground truth information without the need for these locations to be mentioned in the text description.

3.2.3 TIGER

The 2.2 version of the *TIGER* corpus consists of 900,000 tokens, or approximately 50,000 natural language sentences, extracted from the *Frankfurter Rundschau*, a German newspaper (Brants *et al.* 2004).

[27] www.hikr.org/tour/post66901.html

Linguistic parsing, applied to TIGER, consists of a semi-automatically part-of-speech tagging and a syntax parsing. The corpus is generated by the *Institut für Maschnielle Sprachverarbeitung* (*IMS*), the computer linguistic institute at the University of Stuttgart. We use the TIGER corpus, and the part-of-speech tagging in particular, in order to identify common German nouns in corpus data.

3.2.4 DeReKo

DeReKo is the largest reference corpus for the German language. DeReKo consist of more than 5 billion tokens from a variety of language sources, such as fictional, scientific and newspaper texts (Kupietz and Keibel 2009). All texts are part-of-speech tagged and syntax parsed. DeReKo is a product of the *Institute of German Language* (*IDS*), of the *University of Mannheim*. We use the DeReKo in order to normalize frequencies found in text descriptions with expected frequencies in standard German language, as given by DeReKo.

3.3 Elevation Model

As a *digital elevation model* (*DEM*) we use *DHM25*[28] from *Swisstopo*. DHM25 is deduced from the Swiss topographic map 1:25,000, and in particular from vectorized contour lines, point measurements and major breaking lines, such as rivers or lakes. The extraction of contour lines from topographic maps is performed manually. Break lines are defined in a separate, photogrammetric step. In total, 35 to 1600 measurements per km^2 are considered in order to interpolate a continuous elevation grid, with a resolution of 25m. Vertical precision of DHM25 varies with topographic characteristics. In the *Swiss Mittelland*, an extensive plain, and in the *Swiss Jura*, a hilly landscape in northern *Switzerland*, the average precision is 1.5 meters. In the *Alps*, the precision varies between 2 and 3 meters.

[28] www.swisstopo.admin.ch/internet/swisstopo/en/home/products/height/dhm25.html

3.4 Landscape Classification

3.4.1 Arealstatistik

The *Arealstatistik*[29] was first introduced in the early 1980s and is both a land cover and a land use classification for the area of Switzerland. The Arealstatistik is a federal product and part of Swiss constitutional law, which foresees a complete inventory every 12 years. The original motivation for compiling the Arealstatistik was to estimate the areal distribution of cantons and communities in Switzerland. Nowadays, focusing on much finer spatial and semantic granularities, the Arealstatistik is the formal tool for measuring land cover change.

For each grid point on a 100m resolution grid (n points = 4,000,000) the land cover and land use is determined and classified into one of 72 available classes. The classes are defined by experts. This suggests considering the Arealstatistik a top down, or formal taxonomy. The formalism, however, does not meet the requirements of a formal ontology, as specified by Guarino (1998), mostly since relations between classes are not further specified. The only structuring of the 72 classes is a flat hierarchy, consisting of four topics. The topics are *settlement* (n subclasses = 36), *agriculture* (n = 13), *vegetation* (n = 11) and *unproductive areas* (n = 12). The frequency of occurrence is not equally distributed over all 72 classes, as is visualized in Figure 19.

Figure 19. Tag cloud reflecting the frequency of occurrence of the 72 classes of the Arealstatistik in Switzerland.

Two observations can be made, based on Figure 19. Firstly, the frequent classes refer to geographic features of the type forest, meadow and grassland. Settlements and alpine features are underrepresented. For settlements this can be related to the extensive list of available classes (n = 36). Alpine landscapes, on

[29] www.bfs.admin.ch/bfs/portal/de/index/infothek/erhebungen__quellen/blank/blank/arealstatistik/01.html

the other hand, are not in the main focus of the Arealstatistik, as is exemplified in Figure 20, where only three out of 72 classes are useful to describe the high mountain region of *Jungfrau* and *Finsteraarhorn* (n sample points > 10,000).

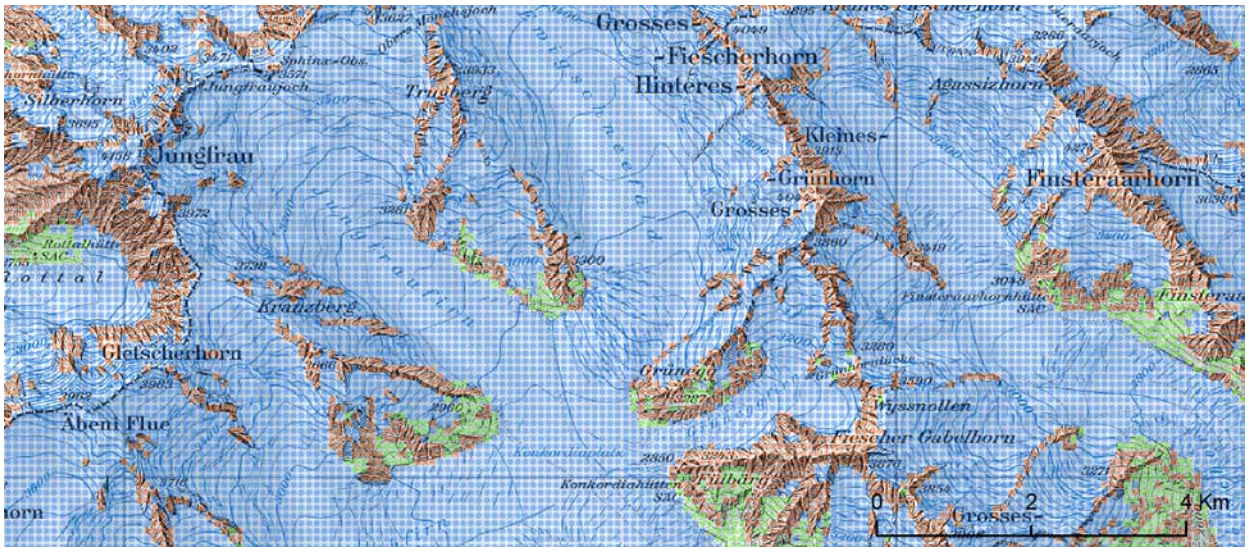


Figure 20. Arealstatistik classification for the *Jungfrau-Finsteraarhorn* region. Three land cover classes are distinguished: Blue = Gletscher, Red = Fels, and Green = Geröll.

The second observation concerns the labels of the classes. Labels often do not reflect everyday terms which are used to describe landscapes and landscape features. A prominent example is the most frequent class *Normalwald*, which would clearly be called *Wald* in standard natural language.

The classification is performed manually and based upon arial image interpretation and stereoscopic representations. Each classification is followed by an automatic evaluation, based on a set of rules, such as the probability of a given neighborhood (e.g. glaciers are more than 100m apart from airports). False classifications are dependent on the frequency of the respective class. For classes that occur at least 1000 times (95% of all classes), the relative error is smaller than 6%.

3.4.2 CORINE

Coordination of information on the environment, i.e. *CORINE*, is a program of the *Commission European*, that started in 1985, and aims to compile information on the state of the environment (Bossard *et al.* 2000). The goal of the CORINE initiative is to gather a consistent body of land cover information, covering the whole of *Europe*, and thus optimally supporting land management, for instance in the context of ongoing changes.

CORINE is a product of manual classification, based on false-color satellite images from *SPOT* and *Landsat*. The manual classification is supported by an automatic quality check. In contrast to the Arealstatistik, which uses a point grid, CORINE stores polygons. The mapping scale is 1:100,000 which corresponds to a horizontal resolution of 250m and smallest units mapped of 250ha.

The classification schema of CORINE is based on a three level hierarchical classification schema. On the first level it distinguishes *artificial surfaces* (subclasses n = 11), *agricultural areas* (n = 11), *forests and seminatural areas* (n = 12), *wetlands* (n = 5) and *water bodies* (n = 7). This results in a final set of 44 land cover classes, all of which are described in great detail to guarantee consistent classification throughout the whole of Europe. In *Switzerland* the compilation of CORINE is organized as a cooperation between *BFS*[30] and *BAFU*[31], both of which are federal institutions.

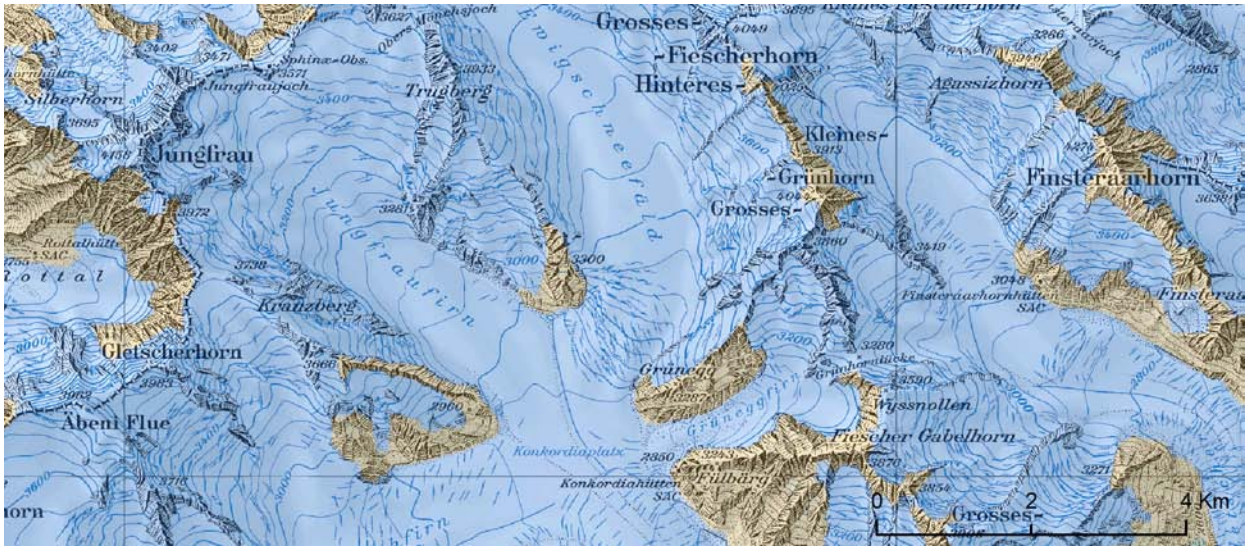


Figure 21. CORINE classification for the *Jungfrau-Finsteraarhorn* region. Two land cover classes are distinguished: blue = Glacier, brown = Bare Rocks.

Figure 21 shows an example of the CORINE classification for the region of *Jungfrau* and *Finsteraarhorn.* It is clear that for high alpine regions only a few classes are available, such that the whole extent represented in Figure 21 is classified as being either *glacier* (blue) or *bare rocks* (brown).

[30] Bundesamt für Statistik, www.bfs.admin.ch
[31] Bundesamt für Umwelt, www.bafu.admin.ch

3.4.3 Swiss Landscape Typology

The *Swiss Landscape Typology* is a classification of *Switzerland* into different planning relevant regions. The typology is applied to all so called *mobile spatial regions* (*ms-region*[32]), as provided by the *BFS* (n = 106). The ms-regions are considered as micro-regions characterized by spatial homogeneity and used for diverse purposes, ranging from scientific investigations to political decision making. For each ms-region a set of 24 criteria are used to generate a grouping. The criteria catalogue incorporates different types of land cover from the Arealstatistik (§3.4.1) and several federal inventories. The result is a grouping of the 106 ms-regions into five homogenous groups, namely *Voralpen*, *Hochgebirge*, *Mittelland*, *Jura* and *warme oder tiefe Lagen* (Figure 22).

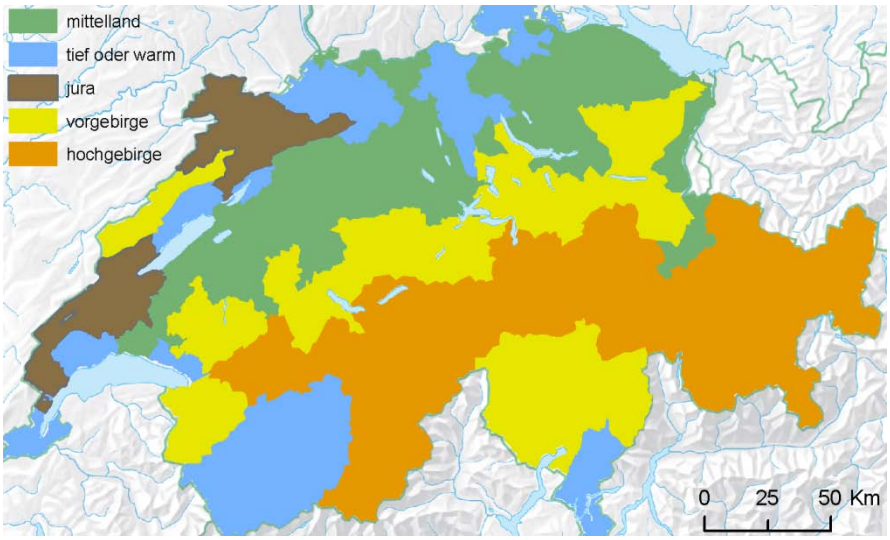


Figure 22. The five *Swiss* landscape types.

[32] www.bfs.admin.ch/bfs/portal/de/index/regionen/11/geo/analyse_regionen/03.html

Chapter 4 Linking Natural Landscape Descriptions to Spatial Footprints

The aim of the first investigation is to link natural landscape descriptions to geospatial footprints. This reflects the initially mentioned role *of* geography in the context of digital humanities and the increasing availability of large compilations of digitized books.

The results of this section consist of a *macro-map* and a spatial index (Figure 23). The macro-map is a spatial representation of the whole corpus as a map. The spatial index, on the other hand, facilitates spatial document retrieval. The spatial index is an important building block for the investigation reported in the next section, which is to automatically compute a spatial folksonomy from descriptions of natural landscapes. The results of this investigation are a contribution to the research gaps RG I and RG II, delineated in §2.3.

Figure 23 is a visualization of the workflow for linking natural landscape descriptions to geospatial footprints.

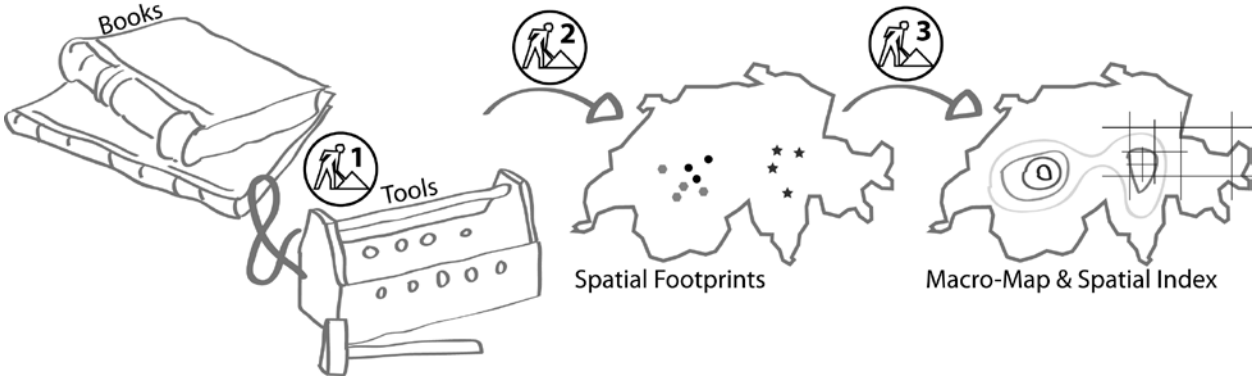


Figure 23. Workflow for linking natural landscape descriptions to geospatial footprints. The work packages are (1) designing and evaluating a toolset, (2) introducing a new approach for geoparsing and (3) computing macro-maps and spatial indexes.

In Figure 23 three tasks are highlighted. Firstly, the design of (1) a toolset, mainly consisting of an approach for measuring geomorphometric similarity. Secondly, we apply this toolset to (2) geoparse a

corpus of natural landscape descriptions, which results in individual spatial footprints of all articles. These footprints are then used for (3) macro-mapping and to build a spatial index[33].

A major task which is not highlighted in the above figure, is the evaluation of the geoparsing approach on the basis of two text corpora. Firstly, we evaluate geoparsing on a corpus consisting of detailed landscape descriptions with the help of expert users. Secondly, we apply geoparsing to a user generated corpus of outdoor activity descriptions, which are associated with rich metadata. For this reason we can perform an extensive automatic evaluation on the performance of our geoparsing algorithm. However, this corpus will not be used for further investigations.

The remainder of this section is a detailed description of all methodological steps needed for (1), (2), (3) and the evaluation, and the representation of all corresponding results.

The work presented in this chapter is covered by following publications:

- Derungs et al. (2011): Disambiguation of Hochmoor descriptions using geomorphometric information.
- Derungs and Purves (2012): Evaluation and application of an approach for comparing toponyms by their geomorphometric characteristics.
- Derungs et al. (2012): Evaluation of a disambiguation approach incorporating geomorphometric characteristics and Euclidean distance, applied to a geographic information retrieval task.
- Derungs and Purves (2013): Detailed description of toponym disambiguation and spatial indexing, however, in the broader context of using text to describe landscapes.
- Palacio et al. (in preparation): An extensive evaluation of a new disambiguation approach for geoparsing natural landscape descriptions, applied to a corpus where each article is associated with metadata.

4.1 Input Data

We mainly use three types of input data for this investigation. Firstly, we geoparse natural landscape descriptions from two corpora, the *Text+Berg* (§3.2.1) and the *HIKR* corpus (§3.2.2). Our application of geoparsing requires the use of gazetteer data and a digital elevation model. As gazetteer data we use the *Swissnames* collection, consisting of more than 150,000 toponym locations in Switzerland (§3.1). In order to gather the geomorphometric characteristics, which are used in our approach of geoparsing, we input the *DHM25* digital elevation model, with a horizontal resolution of 25 meters (§3.3).

[33] Note that our use of the term spatial index is slightly different from its traditional use in GIR, as described in §2.2.1. Details are described in the respective methodology section (§4.2.4).

4.2 Methodology

The methodology follows the workflow sketched in Figure 23 and consists of three tasks. Additionally to these three tasks we describe two approaches for evaluating the outcome of the geoparsing, and a measure for the robustness of the spatial index, computed in task (3).

4.2.1 Geomorphometric Similarity

Geomorphometric similarity is calculated from geomorphometric characteristics, which we gather for a large set of toponym locations, namely all toponyms in the Swissnames gazetteer. Underlying our approach for gathering geomorphometric characteristics of toponyms is the assumption that topography is an important attribute for characterizing landscapes, and landscape features in particular (Smith and Mark 1998). The approach reported in this chapter is described and evaluated in Derungs and Purves (2012). For the evaluation we compared toponyms of different feature types, and could show that there are significant geomorphometric differences between cities, mountains and rivers, and that these differences could not be explained by solely considering spatial proximity. The approach for capturing geomorphometric characteristics incorporates multi-dimensional information from multiple scales.

Geomorphometric characteristics are gathered from values of elevation and slope for a set of three buffer zones[34] (200m, 400m and 2000m) around each toponym location, and thus make a simple association between toponym locations and geomorphometric characteristics (Figure 24). From the distribution of elevations within each buffer zone we store relief (the maximum difference between the elevations of two raster cells within the buffer zone) and standard deviation in elevation (which is related to surface roughness[35]). From the distribution of slopes we retain mean slope and standard deviation. These four measurements (computed for different buffer sizes) are an approximation of the topographic characteristics of landscape features as for instance perceived by humans. The selection of the four topographic measurements, as described above, is not arbitrary. We incorporate variables that are frequently used in geomorphologic classifications (e.g. Iwahashi and Pike 2007) and feature classification algorithms (e.g. Wood 1996).

[34] The different buffer zones are selected such that each feature is characterized through measurements taken on local and regional scales. We assume that the radius of 2000m covers a large share of each features footprint, without incorporating too much of its neighborhood, whereas 200m only represents the *hotspot* of each feature. However, the selection of these three buffers is a pragmatic approximation and will be subject to a critical discussion in the end of this thesis.

[35] Grohmann et al. (2011) consider standard deviation of elevation as one of six possible proxies for surface roughness. In a cross-comparison they found that standard deviation of elevation is particularly suited for explaining roughness on a regional level.

Since both types of measurements are computed for all three buffer zones, we generate 12 attributes that represent the geomorphometric characteristics for each toponym location. The three buffer sizes can be seen as a very simplistic form of a multi-scale analysis. Since measurements taken for the smallest buffer size are again covered in the two larger buffers, these measurements are over represented. This reflects that proximate measurements are considered to be more representative compare to distant measures, which is in accordance with Tobler's (1970) first law of geography (i.e. spatial autocorrelation).

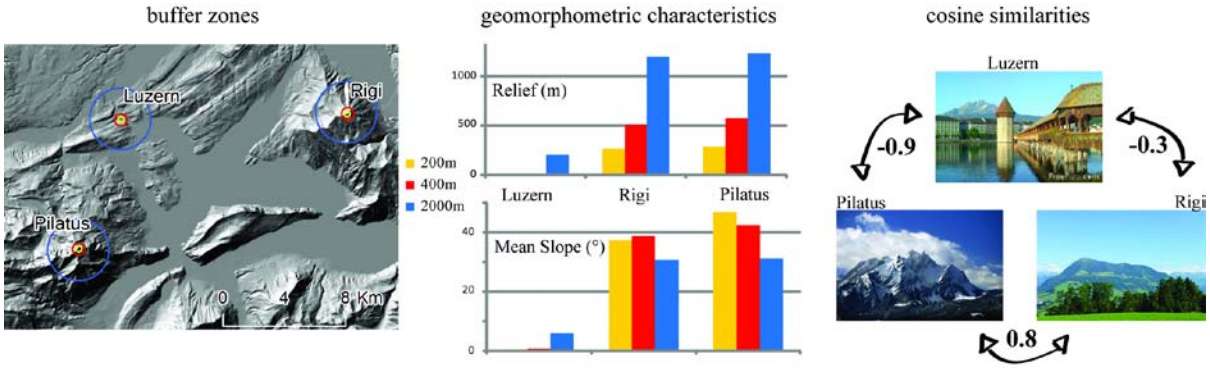


Figure 24. The geomorphometric characteristics (relief and mean slope) computed for three toponym locations and three buffer sizes (yellow, red, blue), with corresponding cosine similarities. (Source Basemap: Swisstopo, Images: www.flickr.com)

Proof of Concept. Figure 24 gives an example of retrieving geomorphometric characteristics for three toponym locations, i.e. *Luzern*, *Rigi* and *Pilatus*. *Luzern* is a Swiss city with approximately 80,000 inhabitants, *Rigi* is an eminence of the type hill or mountain, whereas *Pilatus* is a mountain (Figure 24, right). The differences of the geomorphometric characteristics of the three locations, suggest that *Luzern* has almost no relief and very gentle slopes, whereas *Rigi* and *Pilatus* both have distinct relief, with *Pilatus* being characterized by slightly steeper slopes. These geomorphometric subtleties are visible in the photographs of the three toponyms.

Rigi gives an example why our approach could be considered a (simple) multi-scale attempt for gathering morphometry. The mean slope at *Rigi* is highest for the 400m buffer size, as *Rigi* has the shape of a table mountain, with more gentle slopes in the summit region, followed by steeper slopes at the foothills.

Similarity. The geomorphometric characteristics of toponyms can be represented as feature vectors, such that similarity between toponyms can be computed quantitatively, for example using cosine similarity (e.g. Bayardo *et al.* 2007) (Figure 24, right) (Equation 3).

Equation 3. Cosine similarity calculation for the two vectors A and B, which is computed iterating all dimensions $_i$.

$$cosine\ similarity = \frac{\sum_{i=1}^{n} A_i x\ B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\ x \sqrt{\sum_{i=1}^{n}(B)^2}}$$

Cosine similarity values range from -1 to 1, with -1 indicating inverse trend and 1 congruence. The computation of cosine similarities between individual toponyms and also within or across groups of toponyms of similar feature types, allows for geomorphometric comparisons. Such comparisons could for instance be used to test if all toponyms classified as mountains have comparable geomorphometric characteristics. Our field of application for geomorphometric similarity computations is geoparsing natural landscape descriptions, which is described in the next methodological chapter.

Pseudo Code. In order to allow reproduction of the above methodology we formalized all important steps in pseudo code (Algorithm 1).

Algorithm 1. Pseudo code of three functions for gathering (geomorphCharact()), comparing (geomorphSim()) geomorphometric characteristics of toponyms and generic parts in toponyms.

01: **function geomorphCharact()**
02: computing slope from a 25m DEM → *slope*
03: clipping buffers (200, 400, 2000m) from *slope* and *elevation* surfaces for all toponym locations in Swissnames (n = 156,000)
04: summarizing slope buffers by considering average slope (*aS*) and std of slopes (*stdS*) in each buffer for each location
05: summarizing elevation buffers by considering relative drop (relief) (*rE*) and std of elevations (*stdE*) in each buffer for each location
06: creating geomorphometricCharacteristics by concatenating aS, stdS, rE and stdE for all three buffers for each location
07: **end function**
08: **function geomorphSim()**
09: computing cosine similarity between the geomorphometricCharacteristics vectors
10: **end function**

4.2.2 Geoparsing

The decision to use geomorphometric characteristics in a newly introduced geoparsing approach is, firstly, motivated by literature on the nature of landscape features, as we discussed in the literature review (e.g. Smith and Mark 1998) (§2.1.8). Secondly, we evaluated the suitability of geomorphometric characteristics used in geoparsing in a first case study, published in Derungs et al. (2011). This case study can be considered a simplified test run, since we performed geoparsing on a corpus only describing one particular feature type, namely Hochmoore. As each Hochmoor description is associated with ground truth information, we could show the improvement introduced by using geomorphometric characteristics, over a simple base line geoparsing approach.

Qualitative Description. In theory, geoparsing consists of *toponym lookup* and *toponym disambiguation* (Clough 2005) (§2.2.2). Toponym lookup is performed, using the Swissnames gazetteer (§3.1), and defined as the identification of tokens with similar wording as known toponyms. We call these tokens *potential toponyms*. From these potential toponyms we make a selection of clearly *unambiguous toponyms*, identified as having only one referent location in Swissnames and no generic noun equivalent in standard German language (evaluated using the TIGER corpus: §3.2.3). Unambiguous toponyms are used as anchor points for calculating threshold values for the metrics that are introduced in the following.

Our approach for resolving geo/geo and geo/non-geo ambiguity (§2.2.2) combines two metrics, *Euclidean distance* and, as introduced above, *geomorphometric similarity* (§4.2.1). We therefore call our approach *geometric and geomorphometric disambiguation* (GGD). Euclidean distance is used to compute geometric minimality, which is the resolution of spatial footprints with minimal extent from all referent locations of potential toponyms. Geometric minimality might sometimes be of limited applicability, due to above-average spatial autocorrelation of ambiguous toponyms, as it is indicated by Bunner and Purves (2008) and discussed earlier in this thesis (§2.2.2).

We combine geometric minimality with geomorphometric similarity, as discussed above. Geomorphometric similarity is used to gather the combination of referent locations which are most similar in terms of topographic shape. Both metrics are implemented using threshold values, assuming that geometric or geomorphometric outliers in a document are geo/non-geo ambiguous (introduced in §2.2.2). As indicated above, the thresholds values are gathered from average distances and similarities between all clearly unambiguous toponyms in text.

Geometric proximity and geomorphometric similarity of candidate toponyms are weighted means, computed from neighboring unambiguous toponyms and their respective word distance in text, which is used as an individual weight. Thus, proximate toponyms in text - e.g. toponyms occurring in the same sentence or paragraph - are assumed to be more relevant for approximating the geographic and topographic context of the text. Underlying this assumption is the proximity-similarity heuristic introduced by Tobler (1970), which we think also applies for the use of toponyms in text.

Example. As an example we discuss the application of GGD to one sentence from Text+Berg:

> *Wiederum hatten wir in paar prachtvolle Maitage im Oberland verlebt, hatten Schreckhorn, Agassizhorn und Grosses Fiescherhorn bestiegen.* (from Figure 17)

Oberland, *Schreckhorn*, *Agassizhorn* and *Gross Fiescherhorn* are identified as potential toponyms in the toponym lookup. Of these, *Oberland* is both geo/geo and geo/non-geo ambiguous, as it has four possible

referent locations in Switzerland and is often used as a generic noun in standard German (Figure 25) (toponym ambiguity is described in §2.2.2). The three other potential toponyms are resolved as unambiguous anchor points.

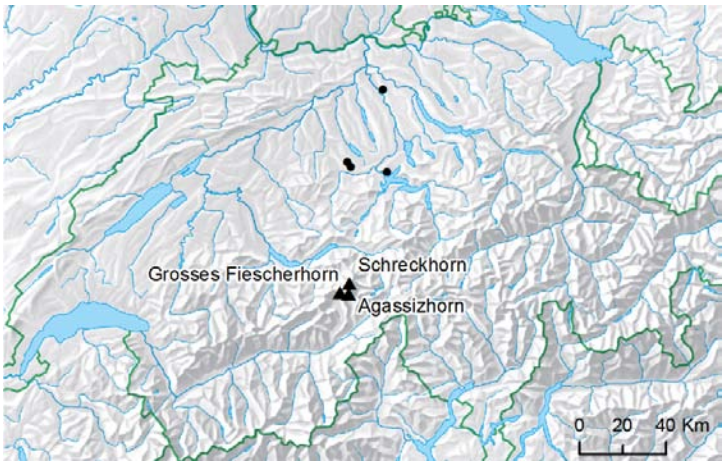


Figure 25. Three mountains (triangles) and the four referent locations of the toponym *Oberland* (dots).

In this case, our method annotates all four referent locations of *Oberland* (black dots) as being geo/non-geo ambiguous, for the reason that all four candidate locations are geomorphometrically unrelated, as well as distant from the three unambiguous mountains, *Schreckhorn*, *Agassizhorn* and *Gross Fiescherhorn*. This is always considered in relation to the geomorphometric similarity and Euclidean proximity shared by the three mountains. The decision for ignoring all four referent locations of *Oberland* stored in our gazetteer is correct. However, *Oberland* indeed refers to a toponym of *vernacular* nature. Thus, the decision for annotating *Oberland* as geo/non-geo ambiguous is incorrect. Vernacular toponyms, by definition, have unofficial status and are thus not stored in Swissnames, which is the official gazetteer, compiled by the Swiss mapping agency Swisstopo.

We applied GGD to two sets of articles (n = 10,000 and 25,000) describing Swiss alpine landscapes, often in terms of outdoor activities. The results of applying GGD are *spatial footprints* for individual articles. In this context, spatial footprints are conceptualized as the set of all toponyms, associated with geographic coordinates, contained in an article. A footprint is thus considered a number of 2D points.

Pseudo Code and Workflow. A detailed list of all important steps for performing *GGD* is given in Table 3. All of these steps are formalized and described in pseudo code in Algorithm 2.

Table 3. Workflow of the GGD geoparsing algorithm.

step	process label	variables
1	toponym recognition	*potTop*
2	tf-idf	*tf-idf*
3	ambiguity	*ambTop*, *unambTop*
4	neighborhood	*neighTop*
5	Euclidean distance fit	*ED_fit*
6	topographic similarity fit	*TS_fit*
7	disambiguation	*ambTop*, *unambTop*

Algorithm 2. Pseudo code of the geoparsing algorithm.

01: **function geoparsing()**

02: **Toponym Recognition:** The text is parsed for terms that have similar wording as toponyms, i.e. *potTop*. As a ground truth set of toponyms we use the Swissnames gazetteer. → *potTop*

03: **TF-IDF:** Term frequency - inverse document frequency values (Equation 1) are calculated for all *potTop*. These values are a proxy for the particularity of terms used in a particular document, compared to the whole corpus. → *tfidf*

04: **Ambiguity:** All *potTop* are evaluated for referent and semantic ambiguity. Referent ambiguity is present if one *potTop* has several referent locations listed in Swissnames. Semantic ambiguity is existent if a *potTop* is tagged as a noun, and not a named entity, in the Tiger corpus. The result is a classification of all *potTop* into ambiguous (*ambTop*) and unambiguaous toponyms (*unambTop*). → *ambTop / unambTop*

all following steps are only calculated for ambTop. UnamTop are resolved as toponyms

05: **Neighborhood:** For each *ambTop* we gather a set of neighboring *unambTop* (*neighTop*). Therefore, only *unambTop* within 200 words distance in text are considered. Each *neighTop* is associated with the *word-count-distance* from the respective *ambTop*. → *neighTop*

06: **Euclidean Distance:** Firstly, we calculate a separate mean Euclidean distance for each referent locations of an *ambTop* and all *neighTop* (*mED_ref*). Secondly, in order to gather a reference value, we calculate the mean Euclidean Distance for all *neighTop* (*mED_neigh*). The minimum *mED_ref*, which is the referent location that is most proximate to the set of *neighTop*, is then related to the *mED_neigh*, the mean distance between all unambiguous neighboring toponyms. This relation is the Euclidean Distance fit (*ED_fit*). *ED_fit* expresses the 'proximity' of the most proximate referent location of an *ambTop*, compared to all *neighTop*. → *ED_fit*

07: **Topographic Similarity:** The exact same procedure as described in step five, however, with the result of calculating a topographic similarity fit (*TS_fit*) as described in Algorithm 1 →*TS_fit*

08: **Disabiguation:** For each *ambTop* we now consider the two values of fit, calculated in step 5 and 6, in order to evaluate if it could be resolved as a toponym. We apply empirical thresholds calculated from cross calculations of fits from all *unambTop*. Thus, we only resolve *ambTop* that either have a *ED_fit* within the 10% best Euclidean distance fits, or, a *TS_fit* within the 5% best topographic similarity fits. The more conservative threshold for *TS_fit* reflects the vagueness comprised in the topographic similarity computation, compared to the straight forward implementation of Euclidean distance.

4.2.3 Macro-Mapping

Macro-mapping is the representation of text as a map and implicitly refers to the additional layer of information which is added to text through the process of mapping. Cooper and Gregory (2011) performed macro-mapping by manually drawing maps from the spatial content of two novels. In our case we are not constrained in terms of number of articles processed and thus create a map of some 10,000 spatial footprints of geoparsed articles. The map is computed from all toponyms resolved from a corpus

and by using a kernel distance for estimating continuous densities. Toponyms are not given equal weights. As an individual weighting function we used the *tf-idf values* (Equation 1). Tf-idf values are a standard measure in IR for approximating the particularity of words in a document, compared to the occurrence of the word in the whole corpus (e.g. Wu *et al.* 2008).

In our case tf-idf values reflect the particularity of the wording of a given toponym in an individual document, compared to its occurrence in the whole corpus. Particular toponyms, i.e. toponyms with high tf-idf values, are considered as being more relevant for a document's footprint and thus given more weight when computing densities.

One of our corpora covers more than a century of landscape descriptions. We thus compute temporal macro-maps for twenty year periods. Temporal macro-maps represent how the spatial focus of the corpus might has changed over time.

In order to highlight particularities in the temporal macro-maps we additionally compute *χ-maps* (e.g. Wood *et al.* 2007). X-maps are spatial representations of *χ-values*, as described in Equation 4 and computed by comparing the density of one temporal macro-map to the density of the whole corpus. The density of a temporal macro-map is the *observed value*, whereas the density of the whole corpus, at the same location, is the *expected value*. X-maps serve as an additional layer of information, such that over- and under-represented regions can be represented separately.

Equation 4. *χ-value* with *Fo* being observed and *Fe* the expected values.

$$\chi - value = \frac{\mathrm{Fo} - \mathrm{Fe}}{\sqrt{\mathrm{Fe}}}$$

The use of χ-maps can be considered a spatial equivalent to the use of tf-idf values for tokens. Both indexes are used to resolve particularities from sample distributions. However, χ-values are computed from metric variables and continuous measurements, whereas tf-idf values are mostly applied to countable variables. A particular property of χ-values, compared to tf-idf values, is that under-representation is explicitly shown, even in cases where no occurrence is measured. In these cases tf-idf values are zero, independent of how surprising a non-occurrence of a certain variable might be.

4.2.4 Spatial Indexing, Ranking and the Adaptive Grid Index

We use two different notions of the term *spatial indexing*. In one sense we use the traditional GIR meaning of the term, where spatial indexing is used to optimize spatial search. As a second meaning, we

compute an *adaptive grid index*, which is a combination of spatial indexing and spatial ranking. The adaptive grid is used to organize all articles in the corpus in a continuous grid with varying resolution. In the following we introduce both notions separately.

Spatial Indexing. As a spatial index we use an *R-Tree*, as implemented in *PostgreSQL*[36] databases. R-Trees were introduced by Guttman (1984) and considered state of the art for indexing point locations in hierarchical, multi-dimensional rectangles, which are allowed to overlap. We apply the R-Tree to index all disambiguated toponym locations, resolved from Text+Berg. Each location is indexed individually, independent from the other toponyms in the same document.

Spatial Ranking. The spatial ranking of documents for spatial queries must consider the spatial index of toponym locations, the association of individual locations to documents and the tf-idf values, as introduced above. The spatial index is used to retrieve a list of documents that contains one or more toponym locations that intersect with a spatial query. The association of locations with documents, and corresponding tf-idf values, is used to rank the list of documents. As a measurement we compute the sum of tf-idf values which is inside the spatial query, as visualized in Figure 26. We call this measurement *spatial relevance* – this does not completely overlap with other meanings of spatial relevance, as for instance described by Raper (2007).

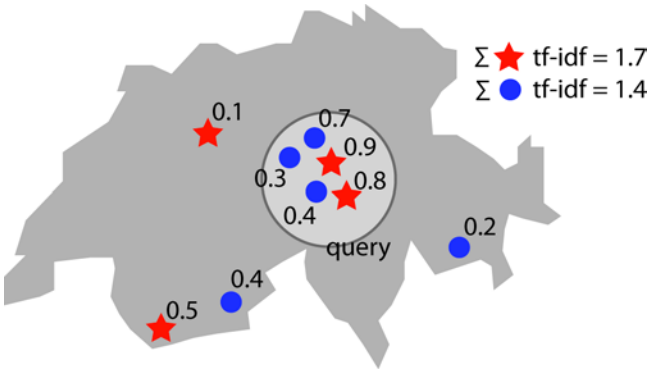


Figure 26. Spatial relevance of two articles (red, blue) based on the sum of tf-idf values of toponyms (stars,circles) inside a spatial query (light grey).

Figure 26 gives an example of the spatial ranking of two articles (*red-star* and *blue-dot*) for a query region (light grey circle). The *red-stars* article has a higher sum of tf-idf values intersecting with the spatial query, compared to the *blue-dots* article, and is thus assumed to be of higher spatial relevance, even though the *blue-dots* article has more toponym locations that intersect with the spatial query region

[36] www.postgresql.org

(n = 3). The tf-idf values are used as a proxy of particularity, or importance, and thus have an effect on the outcome of the ranking.

Figure 27 shows the titles of the top 5 best ranked documents for a quadratic query region (5km) containing the mountain *Matterhorn*.

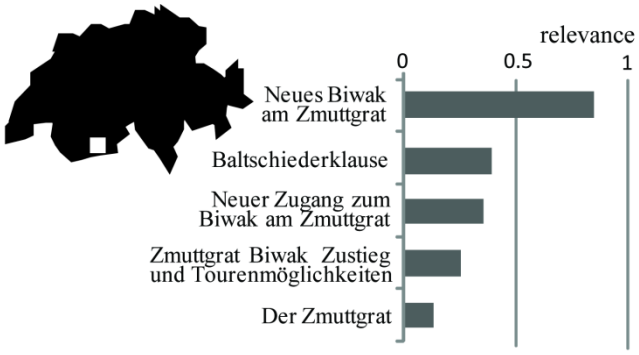


Figure 27. Top five relevant documents for the grid cell containing Matterhorn.

From Figure 27 it is obvious that through ranking we resolve a set of relevant and spatially detailed descriptions. Most descriptions are on ascending the *Matterhorn* by the *Zmuttgrat*, which is a challenging route, following the exposed and steep north-west ridge. The article *Baltschiderklause* is a bias. *Baltschiderklause* is a small mountain hut some 50km distant from *Matterhorn*. The reason that this article is spuriously referenced to the *Matterhorn* region is a problem with digitization. The title *Baltschiederklause* was associated with the wrong text (recent yearbooks have booklet format with several columns). The falsely associated text contains many spatial references that refer to the *Matterhorn* region and is thus considered as one of the top 5 articles.

Adaptive Grid Index. The spatial relevance heuristic can be used to retrieve ranked lists of documents for individual spatial queries, or it can be applied to assign documents to a continuous grid, where each grid cell is associated with a relevance-ranked list of documents, as exemplified in Figure 27.

We develop a grid index where the resolution of a particular grid cell reflects the quantity and quality of articles available for the respective spatial extent. Quantity and quality reflect the sum of spatial relevance values (not normalized) of all documents that are retrieved and ranked for each cell. We therefore use an *adaptive grid* consisting of cells of four resolutions; 40x40km, 20x20km, 10x10km and 5x5km (Figure 28). We thus assume that we cannot retrieve relevant articles from a resolution better than 5km and that the retrieval of documents for extents larger than 40km is not feasible in Switzerland and for this corpus.

The heuristic for computing the adaptive grid is to compute the sum of the spatial relevance values of documents retrieved for each grid of 40km resolution and iteratively double the resolution if above average relevance sums are measured. This process is repeated until for cells with comparably high relevance sums, a maximum resolution of 5km is reached.

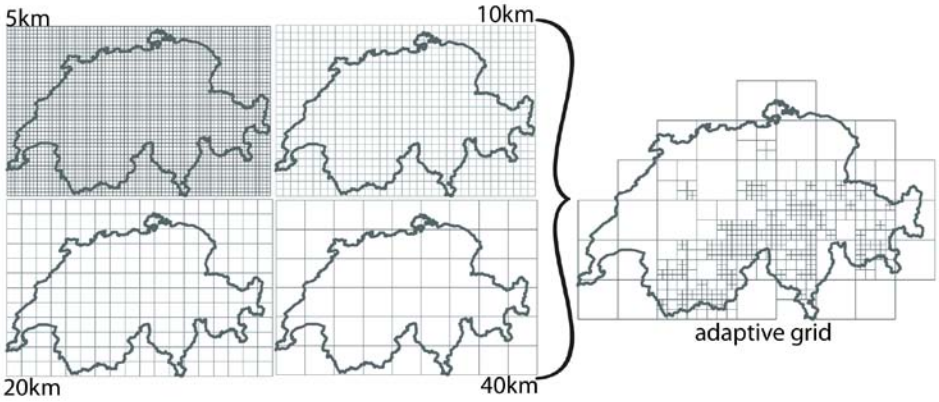


Figure 28. Four continuous grids with the resolutions 5, 10, 20 and 40km.

Our indexing approach is closely related to state of the art *quadtrees* (e.g. Samet 2006). The reason for introducing an own adaption is intrinsic to our data structure, where an individual document often has several referent locations, each associated with an individual weight (tf-idf value). By using the spatial ranking as described above we make sure that the data structure, and in particular the relations between individual toponym locations and documents, is considered in the characteristics of the grid index.

Robustness of Adaptive Grid Index. The adaptive grid index introduces a sharp tessellation of space which is neither intrinsic to the data nor to Swiss landscapes. The tessellation is highly dependent on the resolution(s) of the adaptive grid and the geographic coordinates of the most peripheral referent locations resolved from any of the geoparsed articles. An individual mountain or valley could thus be split into parts, which means that we also split documents that describe this mountain. This might be an artifact of assuming that the earth's surface consists of continuous values, and thus ignoring that the way it is perceived and described is mainly object based (e.g. Smith and Mark 1998). We do not aim to contribute to this research gap, since we have a relatively pragmatic need for a spatial index that is applicable. What we do consider, however, is the effect of the tessellation on the indexed content. Thus, we investigate the change introduced to the index by shifting the location of all grid cells. *Change* is measured as the relative difference of the top 20 ranked documents of each cell after applying 100, 500, 1000 and 2000 meter shifts in different directions.

Small shifts (e.g. 100m) should not have a major effect on the top ranked articles, meaning that the exact location of the grid does not introduce an artifact. Larger shifts (e.g. 2000m), however, are expected to have significant impact on the index, which supports the hypothesis that the descriptions in our corpus are of fine spatial granularity.

4.2.5 Evaluation

Evaluation can take the form of *component*, *system, interaction* or *user centred* evaluation (Mandl 2011) (§2.2.1). We decided to conduct two different user centred evaluations. The reason for conducting user centred evaluations is the lack of gold standard corpora. A gold standard usually consists of annotations of all toponyms occurring in text and the association of the correct referent location. Gold standard evaluation is the simplest way for showing the accuracy of geoparsing. However, since gold standards are rare (and not available on fine spatial granularity), most evaluations are dependent on relevance judgments gained in user studies or from metadata.

In the following we describe two evaluation approaches from applying GGD to two different corpora. Firstly, we describe a user centred evaluation based on the Text+Berg corpus (§3.2.1). This evaluation is crucial since we reuse the results in follow up investigations.

Secondly, we apply GGD to HIKR (§3.2.2). HIKR is associated with rich metadata - not of the type gold standard though - which can be used to conduct an extensive automatic evaluation.

4.2.5.1 Experimental User Evaluation

In order to evaluate the output from geoparsing Text+Berg (i.e. the accuracy of the spatial footprints), we conduct a user experiment and evaluated the improvement of GGD over a simple baseline approach. The user experiment has the form of an information retrieval task. It requires a test collection, such as for instance TREC (Voorhees *et al.* 2005), a set of queries and relevance judgment, obtained by asking users to determine if results are relevant or not. The queries incorporate the spatial dimension only, such that users are presented a spatial extent on a map, together with a set of retrieved documents. The task is to distinguish relevant from irrelevant documents.

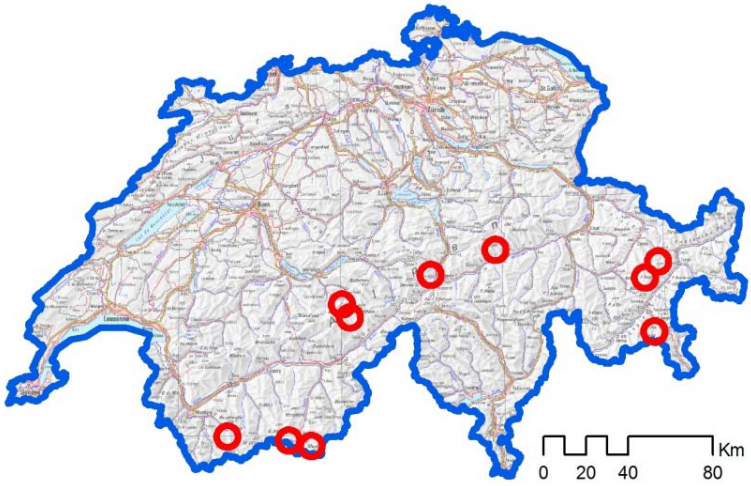


Figure 29. 10 spatial queries for the user centred evaluation.

Relevance and Ranking Judgments. We choose 10 spatial queries – i.e. 5km buffer zones around 10 mountain huts in the *Swiss Alps* - as shown in Figure 29. The 10 query regions are well covered by articles in the corpus. We submitted the 10 queries to 2 retrieval approaches: a simple one that randomly selects a referent location in case of ambiguity (a typical baseline approach where no other knowledge is available, c.f. Clough 2005), and our approach, GGD, as described above. From both approaches we selected the top 5-6 ranked articles for each spatial query and merge them to form a list consisting of at least 9 unique articles (therefore the incorporation of either the top 5 or the top 6 articles, depending on the overlap of the two approaches). The relevance of each of the (at least) 9 articles for each query is judged by 5 participants (i.e. *relevance judgment*). Additionally we asked participants to identify and rank the three most relevant articles for each query (i.e. *ranking judgments*). The ranking judgments allow more detailed interpretations of the performance of the two approaches. Participants were given print outs of all articles for each query and detailed topographic maps in order to better evaluate spatial relevance (1:50,000 and 1:25,000 from Swisstopo).

Local Knowledge. Our test participants were selected by the criteria of being experts in the field of Swiss alpine landscapes. Most of them work in physical geography and have test sites in the *Swiss Alps*. This is important, since labeling documents as being relevant or not for a spatial query is highly dependent on local knowledge (c.f. Purves *et al.* 2007). The dependency on experts has a limiting impact on the number of test participants. We had a total of 12 expert users considered in the evaluation.

4.2.5.2 Metadata Evaluation

From Metadata to Ground Truth. We called this evaluation *metadata evaluation* for the simple reason that it is based on metadata information which is contained in the header of each HIKR article (§3.2.2,

e.g. Figure 18). The metadata is added to HIKR articles by the authors themselves, in order to classify the content of the descriptions. From the metadata we use the activity classification and the way points. The activity classification was used as topical ground truth information, such that the text description of an article classified as *hiking* is assumed to be on *hiking*. The waypoints are used as spatial ground truth information, such that they are assumed to represent important spatial anchor points of the respective description. The activity classification and the way points were both used to index some 25,000 German articles in HIKR. Thus, for each region and topic we can gather a set of articles which we consider relevant ground truth articles for the given spatial and topical specification.

Queries. Each query consists of a topical and a spatial part, i.e. topical and spatial query. An example of a query is *Hiking in Zermatt*, where *hiking* is the topical and *Zermatt* the spatial part, respectively. The spatial part is interpreted as a set of geographic coordinates, centered on *Zermatt*, and the spatial preposition *in* is approximated by separately testing different buffer sizes, i.e. 1, 2, 5, 10km (Figure 30, inset). We use different buffer sizes in order to evaluate the impact of spatial granularity on the retrieval precision.

The different buffers can be associated with different affordances. A buffer of 1km could for instance reflect local information need. Larger buffer sizes, on the other hand, are interesting if one wants to discover a region on foot (5km) or by bike (10km). As topical queries we test all categories used in the HIKR metadata (e.g. hiking, mountain biking, climbing, mountaineering, etc., full list in §3.2.2).

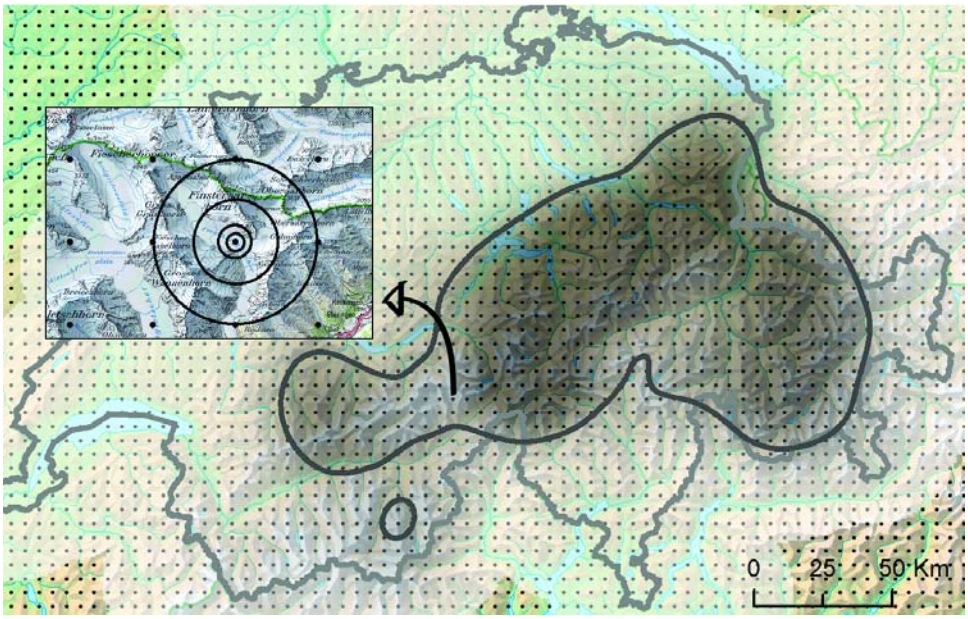


Figure 30. Density of skiing articles in HIKR, with the 20% top density volume as a contour line. Inset: An example of a spatial query and the applied buffer sizes 1, 2, 5 and 10km.

The spatial queries were equally distributed all over Switzerland (5km grid) under, however, the constraint of only testing feasible combinations of spatial and topical queries. For this reason, we firstly compute topic regions, defined as the top 20% volume of the density surface of each topic, as computed from the HIKR metadata. One example is given in Figure 30. Queries on the topic skiing are only combined with spatial queries that refer to locations inside the topic region of skiing.

The result of a particular query consists of the intersection of the results gathered for the spatial and the topical query. The result of the query *Hiking in Zermatt* for instance consists of documents that are on hiking (topically) AND on *Zermatt* (spatially). The nine topics covered in HIKR, combined with all grid points inside the respective topic regions and each tested using 4 buffer sizes, sums up to a total of some 5000 queries. This can be considered a very extensive evaluation.

Approaches. We compare three different approaches for retrieving articles for each query. Firstly, we use the spatial indexing and ranking as resulted from applying GGD, in order to retrieve a set of relevant articles for spatial queries (e.g. Figure 27). In this case, the spatial query is considered a pair of coordinates and the different buffer sizes are used to generate spatial query regions. The topical part of the query is based on tf-idf values computed for all words in the text descriptions of all articles.

The second approach is a simple string baseline (*BL*), where the spatial and the topical query are both considered as strings. For the spatial query this means that a pair of coordinates, which is the original spatial input to the query must first be translated into a toponym. This is realized by resolving the nearest neighbor toponym from Swissnames (§3.1). The ranking is performed using the sum of the tf-idf values from both the toponym and the topic. The different buffer sizes have no impact on the results gathered through BL.

As a third approach we use a spatial query expansion (*SQE*), where the spatial query consist of a list of all Swissnames toponyms that intersect with the particular buffer size (e.g. Fu *et al.* 2005). The list of toponyms is then considered as strings. In combination with the topical query they are used to gather sums of tf-idf values for each article.

Evaluation Protocol. The evaluation is performed by comparing the retrieval results from GGD, BL and SQE, with all relevant ground truth articles for each of the 5000 queries and thus compute precision and recall, as described in §2.2.1.

4.3 Results and Interpretation

For the results and the interpretation we firstly focus on the evaluation of the geoparsing approach. Secondly, we discuss the visualization of all retrieved spatial footprints as macro-maps. Finally, we will represent the adaptive grid index. The adaptive grid index is an important building block of the investigation described in the next chapter (Chapter 5).

4.3.1 Evaluation

We present two sets of evaluation results from applying the GGD approach to two different corpora. The user centered evaluation from applying GGD to Text+Berg is based on 10 spatial queries and judgments gathered from 12 expert users. The number of results is clearly small, but the evaluation task requires local knowledge of the *Swiss Alps*, which adds to the credibility of retrieved judgments but at the same time limits the number of available participants.

The second evaluation is based on metadata, rather than user judgments. On the positive side, this allows for testing a large number of queries, consisting of spatial and topical information. Thus, the metadata evaluation has large spatial coverage and the queries are of fine spatial granularity, which is, to our knowledge, unique in geographic information retrieval. However, the evaluation is based on the assumption that metadata can be treated as ground truth.

The two evaluations cover the same approach applied to different corpora of different coverage and granularity. Consequently, a combined view on both evaluations is complementary such that it informs on general characteristics and the applicability of GGD.

4.3.1.1 User centred Evaluation of Text+Berg

We obtain two types of results from the user centered evaluation, namely relevance and ranking judgments. The relevance of documents retrieved through GGD and a baseline was judged by an expert group (n = 12), familiar with Swiss mountain landscapes. As shown in Figure 31, an average of 82% of the top 5-6 articles gathered for each of the 10 spatial queries were judged to be relevant. This is significantly higher than the 55% gained with a simple baseline approach (t-test: $p<0.05$). A precision (p@5-6) of 82% is relatively high compared to expected precisions as reported in the GIR literature and discussed in the literature review (§2.2.1).

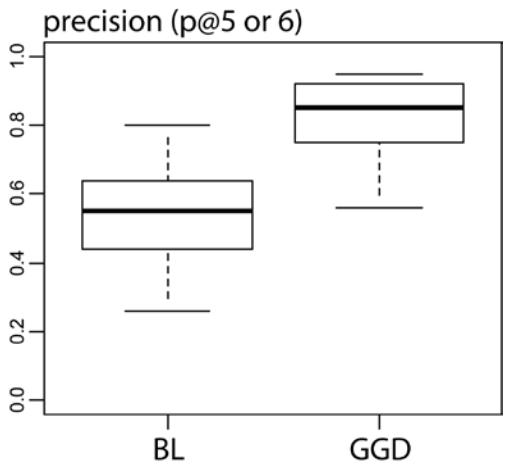


Figure 31. Precision from relevance judgments for the baseline (BL) and GGD disambiguation approaches.

Articles retrieved through GGD tend to be longer than those retrieved with the baseline, whereas baseline articles appear to be focused on the topic when only titles are considered. This is due to the incorporation of sums of tf-idf values for computing spatial relevance (§4.2.4). Articles thus need not be explicitly devoted to only one region. The ranking is good as soon as documents contain relevant spatial descriptions. In general, participants seem to favour longer descriptions, however, for one particular query (*Monte Rosa region*) the titles appear to have strongly influenced relevance judgements (80% base line vs. 56% GGD precision).

Figure 32 contains a summary of the ranking judgments, where participants were asked to identify and rank the three most relevant articles for each query. In the following we use the term *ranked articles* for the ranking introuced by user judgment and *system ranked articles* in order to refer to the ranking which is produced by the two algorithms.

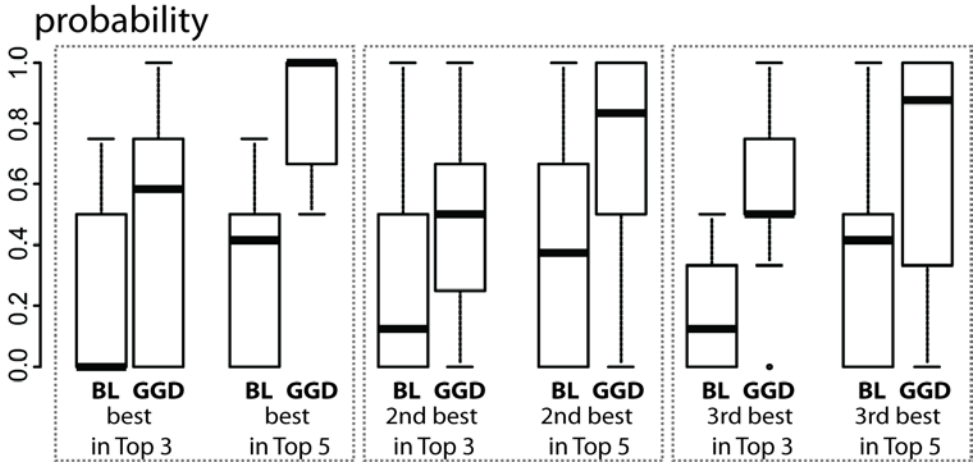


Figure 32. Probabilities based on the ranking judgments, that the best, second best and third best ranked article of a query is listed within the top 3 and top 5 articles, comparing the baseline (BL) and our approach (GGD).

The probability that the three most relevant ranked documents of all queries are listed within the top 3 or top 5 system ranked articles is clearly higher for our approach (GGD), compared to the baseline (BL). The quality of our approach is most obvious when comparing results gathered for the most challenging task, namely the retrieval of the most relevant article (*best*), within the top 3 system ranked articles (left boxplots). Our approach performs with a median probability of 60%, whereas the baseline, in most cases, fails to list the most relevant article within the top 3 system ranked articles.

The relevance and ranking judgments both indicate that GGD clearly outperforms a simple baseline, with a precision value that is relatively high compared to values reported in literature. The high precision is presumably linked to the high availability of relevant documents for the 10 query regions, which are all centered on well-known mountain huts.

4.3.1.2 Metadata Evaluation of HIKR

The HIKR corpus consists of some 25,000 German articles that refer to Switzerland. The evaluation is based on some 5000 queries, mainly covering the *Swiss Alps*, in combination with nine topics, associated with mountain outdoor activities. The exact number of queries depends on the buffer size, since large buffer sizes usually allow to retrieve more relevant articles, compared to small buffers that sometimes have no intersection with any articles.

Figure 33 shows a summary of the precision values for the three approaches, GGD, SQE and BL, and the four different buffer sizes.

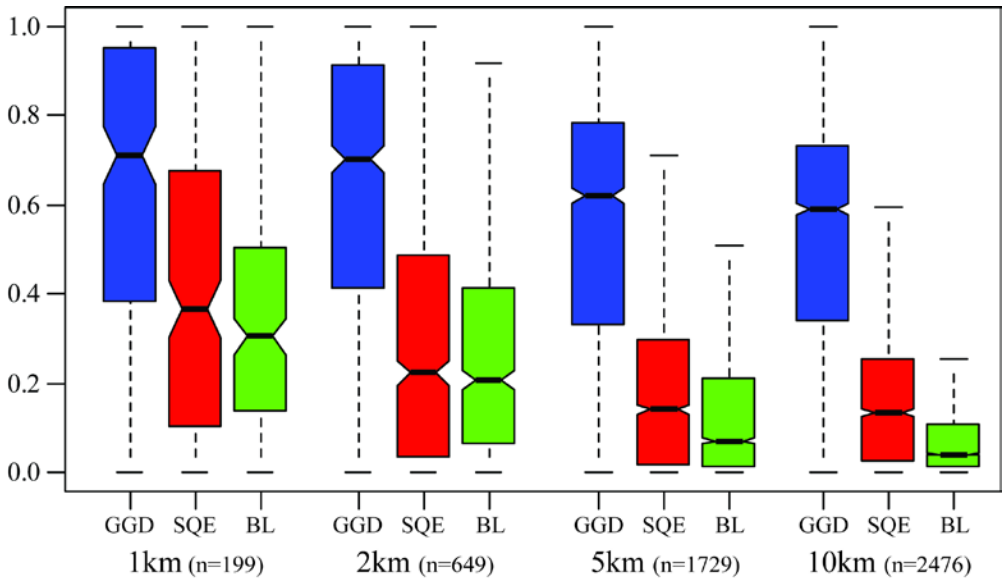


Figure 33. Precision of the three approaches for different buffer sizes.

From the precision values in Figure 33 it is obvious that our approach (GGD) clearly outperforms precisions gained by the other two approaches, which are both string based. This finding is consistent over all tested buffer sizes. The differences are statistically significant, as indicated by the non-overlapping notches of all boxplots. With increasing buffer size the precision of all approaches decreases. The relative difference between the three approaches, however, increases to the benefit of GGD. Thus, our approach seems to be more robust to up scaling, compared to the string based IR approaches.

Since the use of different spatial buffers in GIR can be associated with different affordances, such as local (1 or 2km) or regional interest (5 or 10km), we argue that incorporating geographic information in IR is the only means for retrieving relevant fine grained information on regional level (e.g. 10km buffer: median precision GGD = 0.64 vs. SQE = 0.13 and BL = 0.07).

The differences in precision between the SQE and the BL approaches are less pronounced compared to the precision gained by applying GGD. The BL is always outperformed by the SQE, indicating that expanding the spatial query by a set of local toponyms increases the retrieval precision.

The decrease of precision with increasing buffer size is mainly caused by three effects:

(1) The increasing number of available ground truth articles;

(2) The increasing number of retrieved articles using the two spatially aware approaches;

(3) The increasing number of queries (since we only allow queries for which we retrieve at least one ground truth article, which is more often the case on larger buffer sizes).

This is not quite true for the BL approach. The precision of the BL is robust to effect (2). Thus, the decrease in precision can only be explained by the effects (1) and (3), which describe the increasing numbers of relevant results and queries when buffer sizes increase.

Figure 34 represents the spatial precision of GGD for spatial queries only (similar queries as above, but without the topical parts).

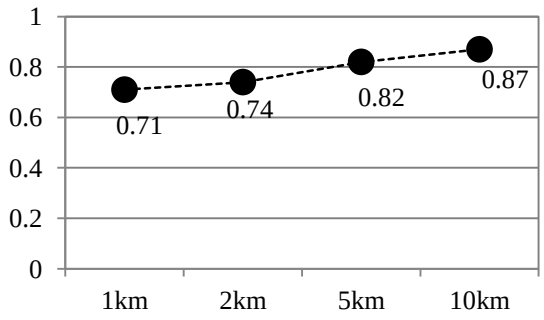


Figure 34. Mean precision of GGD for spatial queries for different buffer sizes.

Spatial precision of GGD increases on larger buffer sizes, indicating that the relative number of relevant results increases as the query region grows. Interestingly, the spatial precision values are comparable to the results gained in the user centred evaluation, where we applying GGD to a different corpus, namely Text+Berg. This indicates that GGD is both generic enough to be applied to different corpora, and detailed, such that it allows GIR with high precision.

In addition to precision, we also compute *recall* (Figure 35). The computation of recall is dependent on knowing which articles from the whole corpus are relevant for each query. This is only given if metadata is available, or only few queries are tested on a small corpus. Most GIR evaluations have not computed recall.

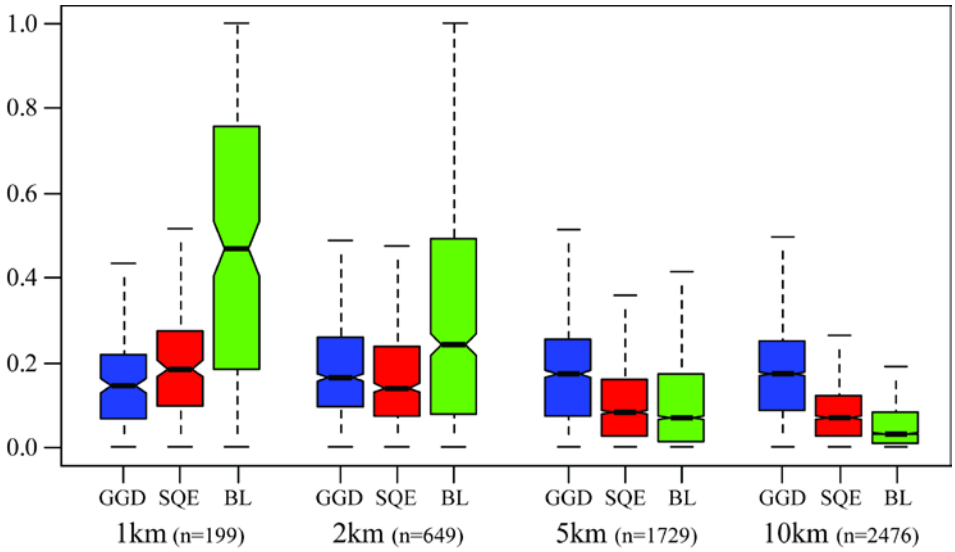


Figure 35. Recall of the three approaches for different buffer sizes.

For small buffer sizes the BL approach outperforms the other approaches. For large buffer sizes the GGD has higher recall compared to the two other approaches. However, recall is low for all approaches and all buffer sizes. This means that for each query only a small share of all available articles can be resolved. The highest recall is gained by the BL approach and for a buffer size of 1km (recall = 0.48), which is due to the relatively large number of articles which are retrieved by searching for a certain topic and toponym in text. The recall of GGD increases slightly with increasing buffer size but is never above 0.21.

The reason for low recall values of GGD are the topical parts of the queries. The recall of spatial queries is considerably higher. This is illustrated in Figure 36.

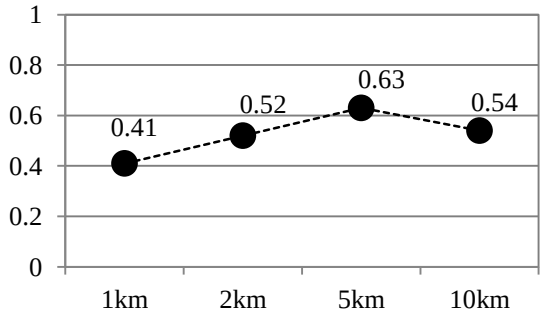


Figure 36. Mean recall of GGD for spatial queries for different buffer sizes.

The recall of GGD for spatial queries is between 41% and 63%. In combination with the precision values for spatial queries (Figure 34) we retrieve a best F_1 accuracy (Equation 2) for the buffer size 5km of some 71%. Compared to literature these are high values.

Based on the literature review of GIR systems, we may conclude that applying GGD to natural landscapes descriptions - which state of the art disambiguation approaches are not well suited for – allows the retrieval of fine spatial granularity information. Thus we could show what in GeoCLEF (e.g. Kornai 2006), the most extensive GIR evaluation initiative, was not obvious; namely that incorporating geographic intelligence in IR can clearly outperform classical IR systems. In SPIRIT (Purves *et al.* 2007) they discovered that GIR can do better than IR, however, only under the condition of incorporating complex spatial relations in the queries, such as directions or distances. In our case we only use the relation *in*, i.e. the simplest spatial relation, and can still show significant improvement.

4.3.2 Macro-mapping

Figure 37 shows a macro-map of the Text+Berg corpus, which is a kernel density map computed from all toponyms grounded in some 10,000 articles and weighted using individual tf-idf values. The map shows isolines indicating the maximal 5%, 10% and 20% volume of the density surface.

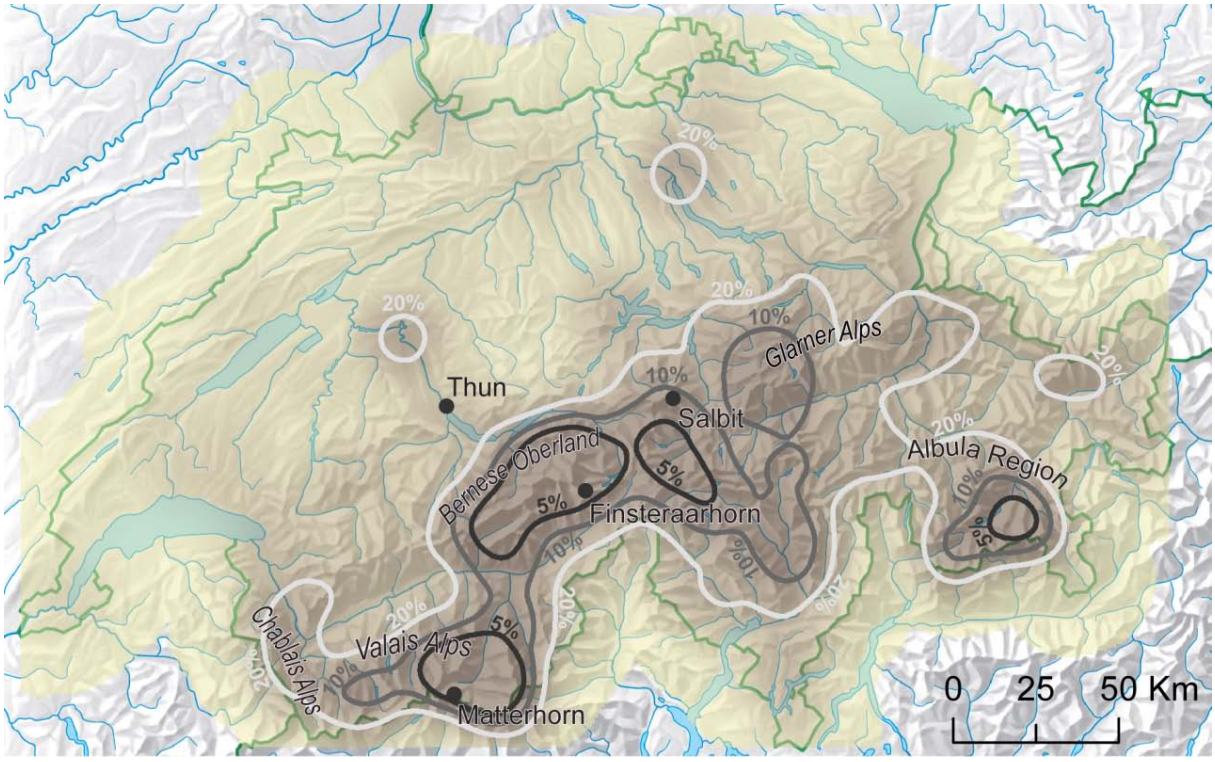


Figure 37. Macro-mapping of Text+Berg, based on a density map from all grounded toponyms in the corpus.

That the core of the corpus lies in the *Swiss Alps* is clearly represented, as is a bias for German speaking regions, with lower values in the Italian part of Switzerland and variations in density in the French speaking parts. The variations in density in the Italian and French speaking parts are most probably not

caused by a language bias of the corpus, as we discussed it in §3.2.1, but by the spatial focus of the topics in Text+Berg. Core areas are found in the *Bernese Oberland* and *Valais Alp*s, crucibles of alpinism in Switzerland where most 4000m peaks are located, with secondary regions such as the *Glarner Alps* and *Albula Region* also visible. Within the 20% most dense areas also the two cities *Bern* and *Zürich* can be identified.

As a means of comparison, we visualize the macro-map of Text+Berg in combination with *topic regions* gathered from German HIKR articles (n = 25,000). Topic regions are computed from user annotations concerning the topic and important way points associated with each HIKR article. In Figure 38 we thus selected three types of activities, namely *hiking* (green), *climbing* (blue) and *mountaineering* (red), and delineate the most dense regions from the associated way points (top 20% density volumes).

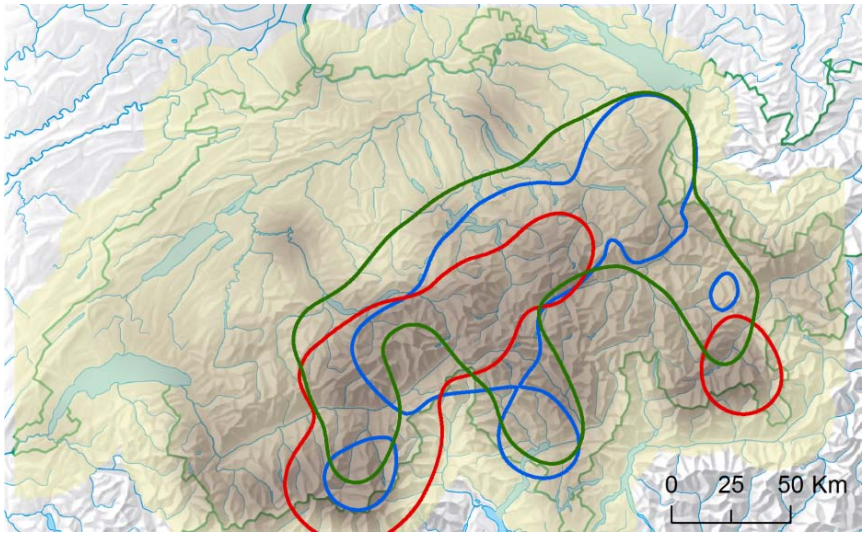


Figure 38. Macro-map of Text+Berg, with activity peaks (top 20% densities) gathered from HIKR entries. Red = Mountaineering, Blue = Climbing, Green = Hiking.

The macro-map shows high overlap with footprints of outdoor activities as described in HIKR. The highest overlap is shown for *mountaineering* (red), indicating that the macro-map represents a footprint of *Swiss* alpine activities and the history of mountaineering in *Switzerland* in particular.

Figure 39 shows density surfaces for 20 year periods between 1860 and 2010.

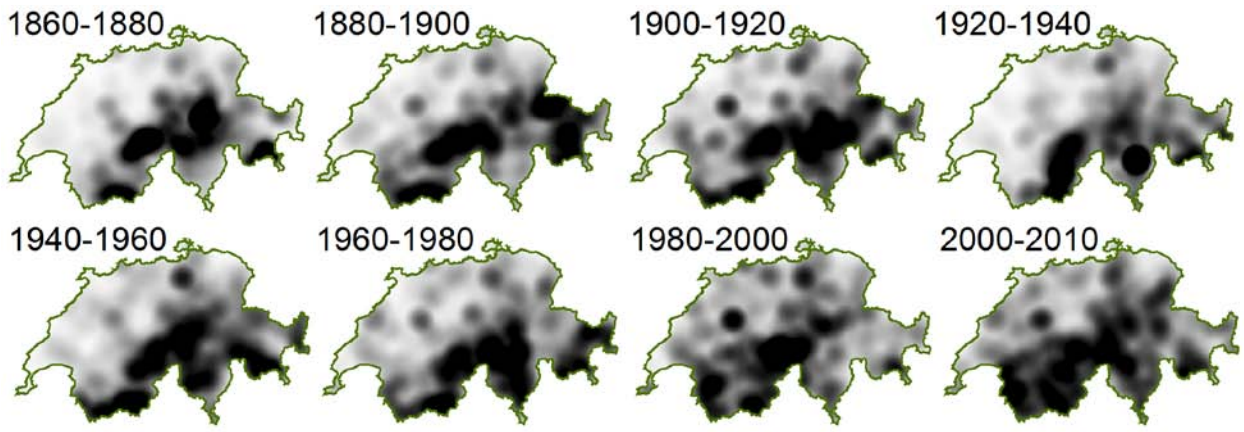


Figure 39. Density surfaces for 20 year periods computed from toponym locations from Text+Berg.

Most temporal macro-maps in Figure 39 have density peaks similar to the macro-map representing the whole corpus (Figure 37). Examples of persistently appearing peaks are the *Bernese Oberland* and the *Valais Alps*. Thus, these regions can be identified as potential target regions for investigations on change of descriptions over time.

The footprints of more recent maps show wider spreads in the spatial distributions. This reflects that over time new places, activities and topics are added to the repertoire of Text+Berg. On the one hand, this overlaps with the growing interest in outdoor activities compared to early decades, where only few people could afford to participate in expeditions in the *Alps* and where Text+Berg almost exclusively reported on mountaineering undertakings. On the other hand, Text+Berg is an edited corpus, published for some 140,000 members of the Swiss Alpine Club. Thus, spatial variation is one important means for keeping the readers interested.

From the temporal macro-maps it is difficult to identify particular events. In order to explicitly visualize particularities and thus allow the detection of events, we represent the temporal maps as χ-maps (Figure 40).

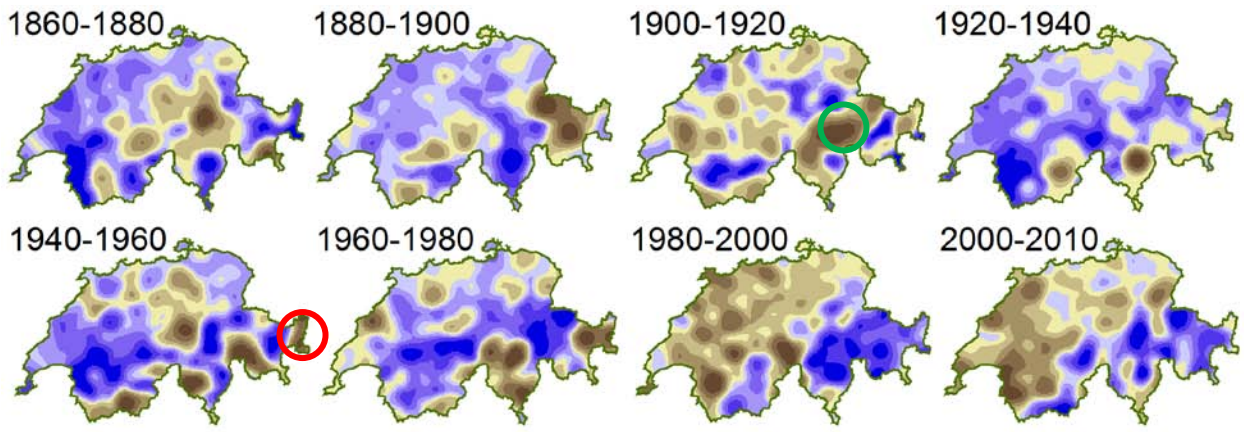


Figure 40. X-maps from density surfaces for 20 year periods computed from toponym locations from Text+Berg. Over-representation is visualized in brown color, blue color indicates under-representation. Similar color values across maps do not necessarily indicate similar χ-values.

X-maps are well suited to highlight regions which are over- (brown color) and under-represented (blue color) with densities. In the χ-maps in Figure 40 we can, for instance, identify over-representation of densities in the *Albula Region* around 1900 (green circle, map 1900-1920), which co-occurs with the opening of the railway. Another example in eastern Switzerland is the opening of the *Swiss National Park* to the public in the 1920s (red circle, map 1940-1960).

An extension to the macro-maps as presented above would be a dynamic linking between the density surface and the underlying descriptions. This would clearly facilitate the detection of topics and events and thus support the readability of large corpus data, which is often too comprehensive for close reading. A similar feature is implemented in the *Google Books Ngram Viewer*[37] (as discussed in the introduction), where temporal footprints of term frequencies are associated with the *responsible* documents.

4.3.3 Adaptive Spatial Grid Index

The adaptive grid shows high resolution for regions described by a large number of documents and in great level of detail. Level of detail is approximated by the sum of tf-idf values of toponyms referring to the respective grid cell. Thus, a description is considered to be detailed if it contains toponyms that are not or only rarely used in other descriptions. The spatial resolution of the adaptive grid clearly correlates with core regions identified in the macro-map of Text+Berg (Figure 37).

Figure 41 is a visualization of the adaptive grid index. Each cell is associated with a relevance ranked list of documents from Text+Berg, where the document relevance is computed from the sum of tf-idf values

[37] books.google.com/ngrams

of disambiguated toponyms that intersect with each cell. Thus, the term *index* is slightly misleading, since in an IR context this would be considered a combination of indexing and relevance ranking.

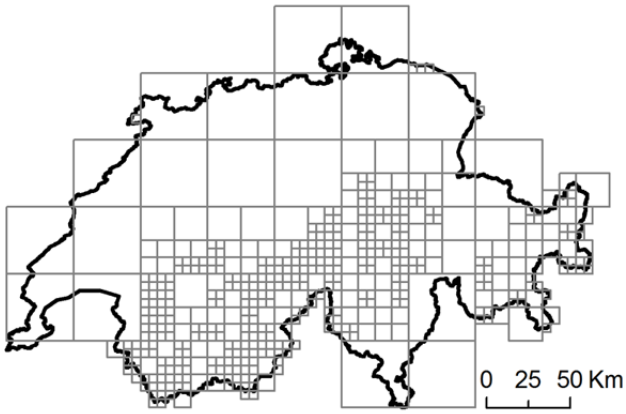


Figure 41. Adaptive spatial grid index computed from spatial footprints.

The grid index is a combination of four predefined grid resolutions, i.e. 40km, 20km, 10km and 5km. We decided for the maximum resolution of 5km to capture the content of descriptions. This approximately covers the footprint of a mountain, several hours of hiking, climbing or mountaineering. From 5km we incrementally halve the resolution up to a minimum resolution of 40km, which is clearly too coarse for individual descriptions in Text+Berg. We introduced a heuristic for combining the four resolutions and compute an adaptive grid that is closely related to the quadtree index. The reason for introducing this indexing approach relates to our particular data structure, where one document is represented by a set of toponym locations, each associated with an individual weight (i.e. tf-idf value).

The exact position of grid lines introduces some boundary effects. Switzerland is thus tessellated into *arbitrary* landscape *units*, such that individual geographic features, such as mountains or valleys, are fragmented or aggregated. In order to control for boundary effects we measured the change in the top ranked documents, introduced by spatially shifting the adaptive grid for some 100, 500, 1000 and 2000 meters (Figure 42).

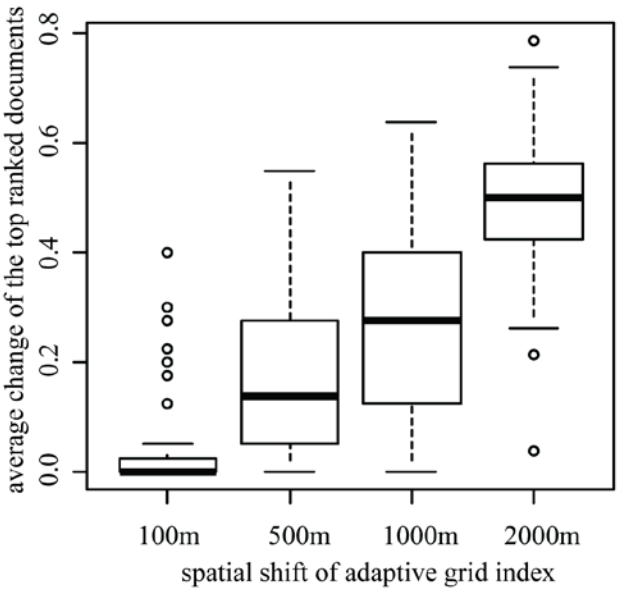


Figure 42. Relative change in the lists of top 20 ranked documents averaged over all grid cells.

A 2000 meter shift of the adaptive grid causes a median of 50% of the top 20 ranked documents to differ. A shift of 100 meters, on the other hand, has almost no impact on the lists of top ranked documents (median: 0%). The spatial index only varies with large spatial shift, which is an expected and desired behavior. It allows us to argue that the spatial granularity of the underlying descriptions is of great level of detail and that the exact location of the adaptive grid is not too critical.

The spatial representation of *change through shift* shows that mainly regions represented with high resolution are prone to change (Figure 43). Change is not spatially autocorrelated, such that no larger region seems to be particularly vulnerable to small spatial shifts. Large change also does not affect grid cells containing, or splitting, prominent mountains. This is important, since in follow up investigations we will focus on some of these cells and analyze their description. It is thus crucial that the description do not significantly change if the adaptive grid is shifted by only some 100 meters. A shift of 2000 meters, on the other hand, has strong effect on almost all cells in the *Swiss Alps* (Figure 43, right).

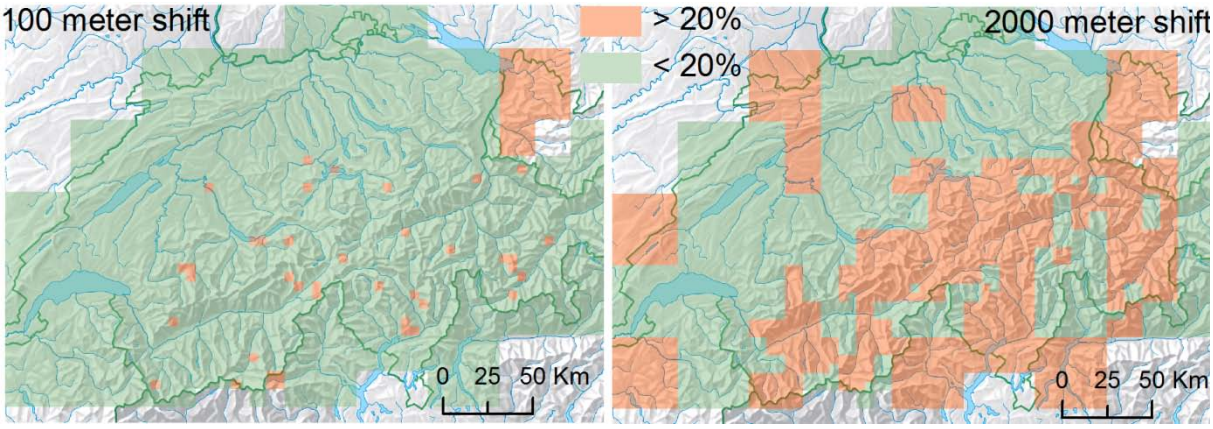


Figure 43. Change (<20% and >20%) introduced to document rankings through spatial shift (100 and 2000 meters).

This indicates that many of the descriptions associated with these high resolution regions are of fine spatial granularity, such that a 2000 meter shift has significant impact on cell contents. This again is an important finding for follow up investigations, since there we argue that based on the adaptive grid index, we can gather detailed local landscape information.

Chapter 5 Extracting Landscape Information from Georeferenced Descriptions

The aim of this investigation is to compute a spatial folksonomy from natural landscape descriptions. Thus, a large compilation of landscape descriptions is investigated for information that is important for contributing to fundamental geographic research questions, such as information on how people describe their local environment in everyday encounters. We called this the role *for* geography in the introduction.

The spatial folksonomy is a vocabulary of natural features, which reflects local subtleties and variation in landscape descriptions in *Switzerland*. It is local as it is retrieved from spatially indexed landscape descriptions (i.e. the adaptive grid index which resulted from the previous investigation as represented in Figure 41).

Figure 44 is a sketch of how we computed the spatial folksonomy.

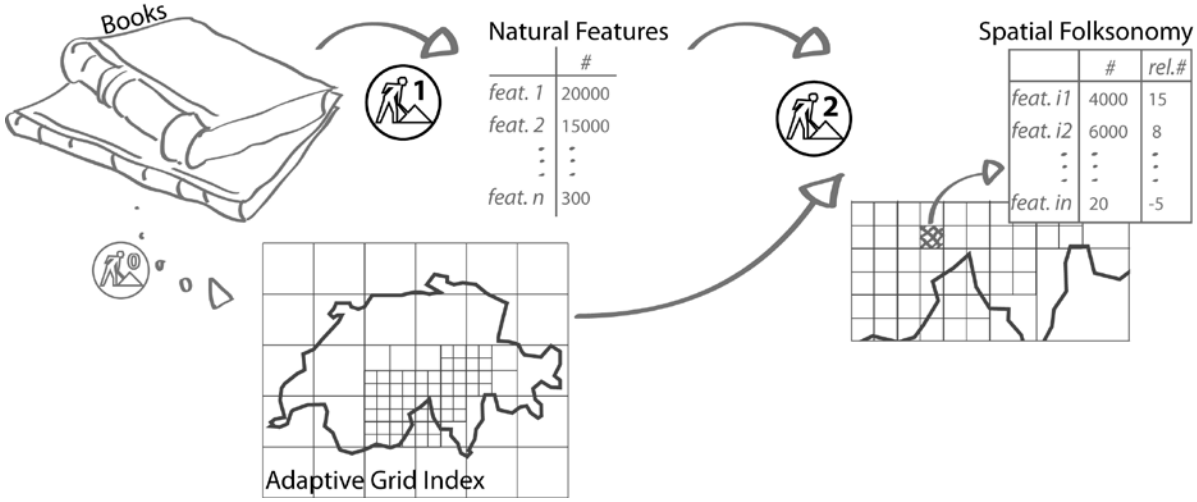


Figure 44. Workflow for computing the spatial folksonomy from natural landscape descriptions. The work packages are: (1) annotating a set of natural features occurring in text, and (2) the computation of a spatial folksonomy, from combining the (0) adaptive grid index, generated in the previous investigation, and the list of natural features.

In Figure 44 two tasks are highlighted. Generating the adaptive grid index (0) is kept in grey, indicating that it is already covered in the previous investigation (Chapter 4). The first task (1) is to annotate

frequently occurring natural features from a corpus on natural landscape descriptions. In a second step (2), we use this vocabulary of natural features and combine it with the adaptive grid index, in order to compute the spatial folksonomy.

The first focus of this chapter is on describing all methodologies needed for (1) annotating natural features and (2) computing a spatial folksonomy.

In the following, we will firstly use the spatial folksonomy in order to describe and compare different landscapes. Secondly, we will evaluate and contextualize the content of the spatial folksonomy by linking it to formalized land cover classifications. Thus, we aim at finding answers to questions such as *How different is the description of Matterhorn from the description of Uetliberg?* and *How can formalized land cover classifications benefit from information extracted from natural landscape descriptions?*

The work presented in this chapter is covered in the following publication:

- Derungs and Purves (2013): This publication is on, firstly, linking a historic corpus of landscape descriptions to space and, secondly, gathering geographic information that represents local landscape descriptions.
- Derungs and Purves (in preparation): A comparison of local geographic landscape information as retrieved from different corpora. Additionally, we compare the geographic information in the spatial folksonomy with land cover classifications and thus conclude on important differences and potential synergies.

5.1 Input Data

We use the natural landscape descriptions from the Text+Berg corpus (§3.2.1), which we received in a preprocessed format, consisting of a part-of-speech tagging, which is provided by the computer linguistic lab of the University of Zurich (Sennrich *et al.* 2009). Additionally, we use the adaptive grid index, which we computed in the previous investigation (Chapter 4). Both input data are combined in order to compute the spatial folksonomy.

In the following part of this investigation we use three official landscape classifications, the *Swiss Arealstatistik* (§3.4.1), the *European CORINE* (§3.4.2) and a *Swiss* landscape typology (§3.4.3). These classifications are used in order to compare their contents to the information contained in the spatial folksonomy.

5.2 Methodology

In this chapter we discuss the methodology, firstly, to annotate a vocabulary of natural features frequently occurring in landscaped descriptions. Secondly, we retrieve frequencies of these natural features from the georeferenced descriptions that resulted from the previous investigation. We call this the spatial folksonomy. Thirdly, we describe how the spatial folksonomy can be used for qualitative and quantitative comparisons, and how the spatial folksonomy can be linked to official land cover classifications.

5.2.1 Natural Feature Annotation

The aim of this task is to resolve a vocabulary of terms from a text corpus that is frequently used to refer to natural landscape. We call these terms *natural features*. This vocabulary will then be used to analyze descriptions of landscapes. Analyzing landscapes through investigating landscape features reflects the notion of landscape as a whole consisting of parts (Naveh and Lieberman 1984). Apart from this theoretical motivation for focusing our investigation on landscape terms only, we identified reasons for focusing on a controlled vocabulary in other work with user generate content (Purves *et al.* 2011). The decision and consequences of only using landscape terms for analyzing landscapes, and not incorporating all terms used in the descriptions, will be discussed in the end of this thesis.

Natural features are, according to Smith and Mark (2003), almost exclusively treated as objects in folk disciplines and represented as nouns in natural language (c.f. Nelson *et al.* 1993). We thus concentrate on identifying nouns within our corpus which refer to natural features. Nouns are identified in a preprocessing task, where a *hybrid tagger*, combined with a *rule-based* and *probabilistic* heuristic is applied to the corpus (Sennrich *et al.* 2009). This task, which is state of the art *linguistic parsing* or *part-of-speech tagging* (POS) is performed by the computer linguists at University of Zurich.

We distinguish natural features from all other types of nouns (e.g. proper names or artificial features) by performing a *manual annotation* task (e.g. Blaylock *et al.* 2009). There, human annotators explore a list of frequent nouns from the corpus and, by applying a set of four rules, identify natural features. We consider this a state of the art approach for selecting a specific group of terms from all available terms in a corpus, as for instance described by Purves et al. (2011). The reason for conducting the annotation task is to gather a list of terms that explicitly refer to natural landscapes. Later, these terms are central for deducing information from landscape descriptions that is comprehensible to a human interpreter. According to the four annotation rules natural features are:

1. generic rather than specific (e.g. mountain not Matterhorn);
2. natural not artificial (e.g. stream not hut);
3. objects rather than activities (e.g. path rather than ascend) and
4. a perceivable object in a landscape not merely a phenomena or qualities (e.g. glacier or snowfield not ice or snow).

Clearly there are a number of boundary or vague cases which are important to distinguish. For example, a *meadow* appears to many individuals to be a natural feature, but is in fact part of a maintained landscape. Thus, our annotation was carried out by four individual annotators, all German native speakers and all furnished with a more detailed description of the rules set out above. The original rules, as given to the annotators are shown in Appendix A. The annotators worked through randomized lists of the 1500 most frequent nouns in Text+Berg, and identified those that they considered natural features according to the rules given. Only nouns classified by three or more of our annotators were retained in the final list of natural features.[38]

5.2.2 Spatial Folksonomy

A standard approach to analyze natural language documents is the so-called *bag of words* approach that often uses *inverted file structures* (Chowdhury 2010) (Figure 45). Inverted file structures only consider term frequencies, instead of complete syntax and context information. We decided to design the spatial folksonomy to consist of inverted files that only contain natural features, as resolved in the previous described annotation task. Figure 45 shows a virtual example of the inverted file of nouns and natural features deduced from a sentence from the exploration of the Grand Canyon.

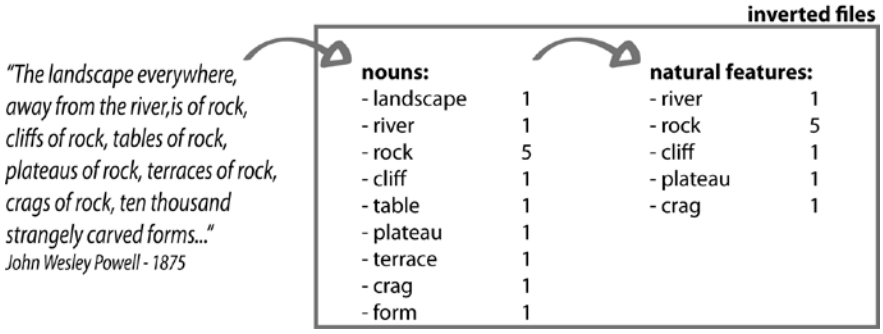


Figure 45. Inverted file consisting of nouns (left) and natural features (right) from a sample sentence.

[38] Some of the vague cases are not intrinsically vague but vague since the nouns, as represented in lists, are shown to the annotators in an out of context situation. It was thus often difficult to identify the true meaning of a noun, in particular if several meanings are available (i.e. ambiguity). We discuss this limitation and potential improvements in the discussion chapter of this thesis.

It is clear from the given example that the inverted file of natural features is only representative for descriptions of natural landscapes. As mentioned above, inverted files are based on term frequencies. Term frequency distributions in language typically follow *Zipf's law* (Zipf 1935), that is to say frequency of terms is inversely proportional to their rank (e.g. Figure 46 – word frequencies in *The Simpsons*). Thus, natural features that have similar frequency ranking can still have very different frequency counts. This influences the statistical analysis of inverted files, since the ranking of natural features in inverted files can be robust, even if the frequency values show pronounced variation. Thus, rank order statistics, such as *Mann-Whitney U* or *Kruskal-Wallis* tests often fail in assessing differences between inverted files. The described effect is particularly pronounced for frequent terms.

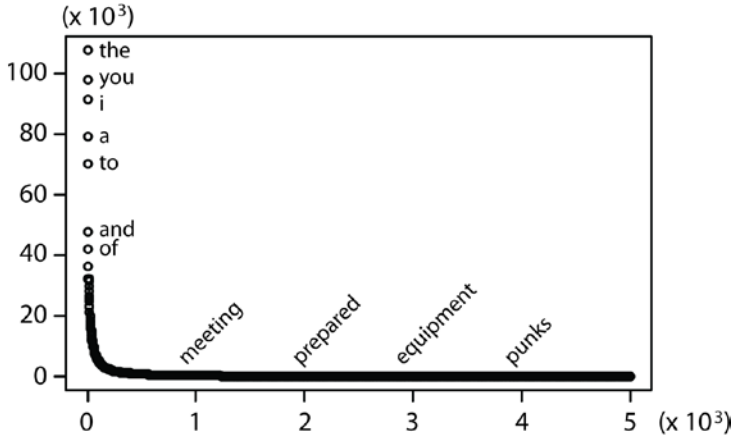


Figure 46. Zipf frequency distribution of the 5000 most frequently used terms in *The Simpsons* (Source: pastebin.com/anKcMdvk).

The example of the 5000 most frequent words used in *The Simpsons* (Figure 46) shows a clear Zipfian distribution. It is obvious that many of the most frequently used terms are not particular, such as *the*, *you* and *I*, whereas more particular words occur in the long tail of the term distribution (e.g. *punks*).

In order to correct for the influence of Zipf distributions, and to detect more fine granular variations between inverted files we rely on normalized frequencies, where the frequency of a term in a document is normalized by the frequency of the term in the whole corpus. Terms that are frequent in a document and in the corpus are considered less important than terms common in a document but rare in the corpus. There are numerous measures normalizing frequency counts. We use *tf-idf* values (Equation 1, p.33), a standard measure in information retrieval that has already been applied in other studies for ranking spatial occurrences of terms (e.g. Rattenbury and Naaman 2009).

The spatial folksonomy is computed from inverted files and tf-idf values of natural features for each grid cell of the adaptive grid index (§4.3.3). The step wise process for computing the spatial folksonomy is described in the following list and sketched in Figure 47:

1. Iteratively, for each grid cell of the adaptive grid index, we retrieve sets of spatially relevant documents, as described in §4.2.3 and in Figure 27.
2. The list of documents is transferred into an inverted file by analyzing each document for the frequency counts of natural features (Figure 45).
3. We compute tf-idf values from the frequency counts of natural features within a grid cell, and information on natural feature frequencies in the whole corpus (Equation 1, p.33).
4. The result of this process is a ranked list of natural features for each individual grid cell of the adaptive grid. We call it a *spatial folksonomy*.

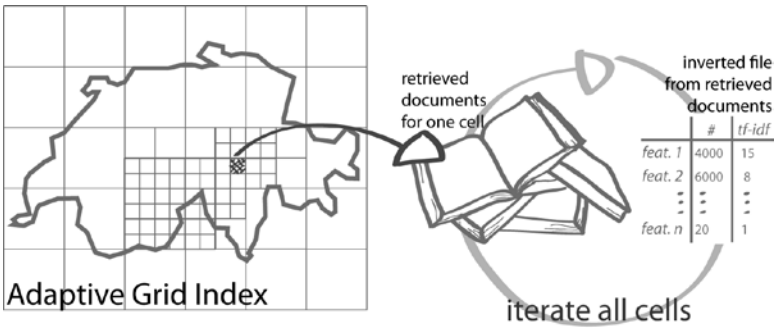


Figure 47. Computing the spatial folksonomy from documents indexed in the adaptive grid.

As we showed in the literature review, folksonomy is characterized by the input data, rather than by a methodological paradigm, as is for instance the case with ontology (§2.1.7.1). We use the term spatial folksonomy for emphasizing that we gather information from text documents that are written for the purpose of describing an outdoor activity or a landscape. Such information is comparable to, for instance, tags used to describe an image or a video. Tags are often incorporated in folksonomies, since they reflect human sourced concepts and descriptions.

5.2.3 Comparing Regions and Natural Features for their descriptions

The spatial folksonomy can be considered a matrix populated by tf-idf values of natural features (Figure 48, columns) for each cell of the adaptive grid (rows). From this matrix we can either extract individual vectors for each (a) natural feature or for each (b) cell.

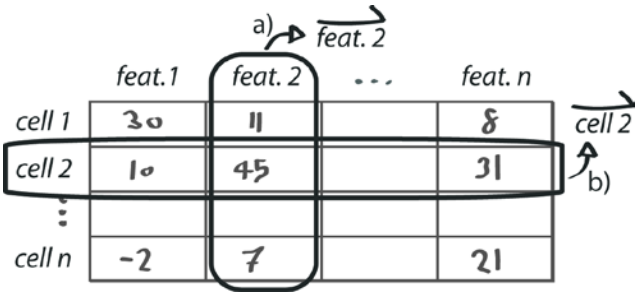


Figure 48. Spatial folksonomy as a matrix, consisting of natural feature (a) and cell vectors (b).

Both types of vectors are numeric and can thus be compared by calculating cosine similarities (as discussed in §4.2.1) or they can be grouped, using a clustering algorithm. We use a simple and well-known clustering algorithm, namely K-means, which is considered a robust baseline for automatically identifying *k* groups defined by most similar vector values (Faber 1994).

Comparing natural feature vectors allow for answering questions such as *In what way is the (spatial) use of the term Berg different from the term Gipfel?*, whereas the comparison of cell vectors allows for comparing different regions for similar descriptions, which is reflected in the question *How different is the description of Finsteraarhorn from the description of Uetliberg?*. We will focus on results gained from answering questions of the latter type, by comparing the local information associated with different regions. In order to represent two diverse regions we focus on *Finsteraarhorn* and *Uetliberg*, both completely covered by individual cells in the adaptive grid (Figure 49). *Uetliberg* is a hill in the north-eastern part of *Switzerland*, neighboring the city of *Zurich*. *Finsteraarhorn* is a prominent mountain in the *Bernese Alps*. Obviously, the two features are located in different parts of *Switzerland*, characterized by different topographies.

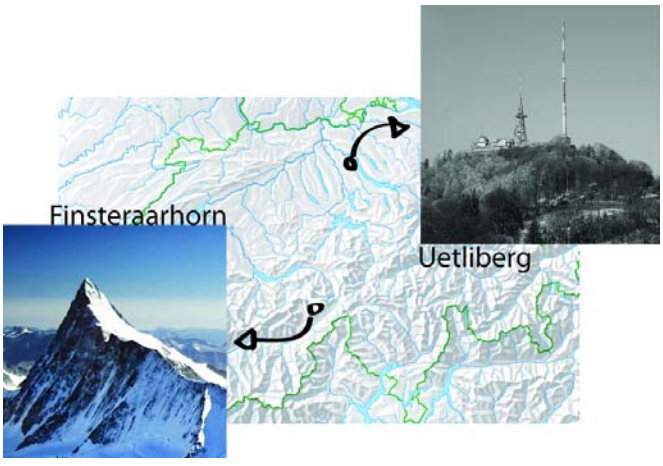


Figure 49. Finsteraarhorn and Uetliberg.

For both cells - containing *Finsteraarhorn* and *Uetliberg* - we compute the cosine similarities to all other cells of the adaptive grid, using all cell vectors. Thus, we generate two maps that show continuous landscape (description) similarities for *Uetliberg* and *Finsteraarhorn*. We call these maps *landscape similarity maps*.

5.2.3.1 Explaining Variation in Landscape Descriptions

In order to put the landscape similarity maps into context we compare them with geomorphometric characteristics, such that we can investigate if variations in descriptions can be explained through geomorphometric variation. As an approximation of the geomorphometry of each cell of the spatial folksonomy we use the relative distribution of the geomorphologic classes, resulting from the algorithm introduced by Iwahashi and Pike (2007) (Figure 7, p.35). Thus, each cell of the spatial folksonomy is associated with the relative distribution of the 16 geomorphologic classes that relate to slope, curvature and texture. These 16 values can be considered numeric vectors, similar to the frequency distribution of natural features, and can thus be used for computing similarities between all cells of the spatial folksonomy. The maps generated from the geomorphologic similarities we call *geomorphologic similarity maps*.

5.2.3.2 Computing Landscapes from Landscape Descriptions

Clustering, applied to cell vectors, can be used to resolve *k* groups (as cluster size in K-means) of similarly described landscapes all across Switzerland. Accordingly, clustering can be used to answer questions such as: *What different types of landscapes can be identified in Switzerland, in terms of their description?* The product can be discussed in the light of other initiatives of automatically generating landscape typologies (e.g. Van Eetvelde and Antrop 2009). In order to evaluate if the landscape typologies correspond to well-known types of landscapes in *Switzerland*, we compare the clustering results with an official *Swiss* landscape typology, which knows five landscape types (§3.4.3). High similarity between the clustering result and the official landscape typology would suggest that we can reproduce meaningful spatial entities by going from georeferenced text to spatial descriptions in our folksonomy.

5.2.4 Spatial Folksonomy and Land Cover Classifications

How can land cover classifications benefit from landscape descriptions? Land cover classifications are often compiled in order to quantify the earth's surface and to monitor change of landscape over time. Therefore, the taxonomy of land cover classes must be consistent over space and time, such that one class

is applied according to the same standards at different locations, in consecutive inventory years and across annotators. The interoperability of land cover classifications is guaranteed by using formal application rules that clearly define each class, its correct application and how it is distinct from similar classes. Another particularity of land cover classifications is that the taxonomy is designed for a particular purpose. Thus, it is often defined by experts and contains a great level of detail for places and topics of interest. In peripheral regions the focus is more on efficient classification.

The application rules of land cover classifications have three consequences. Firstly, the taxonomies are not equally well suited for classifying all types of landscapes. Secondly, the classification rules are the same everywhere, such that local subtleties in landscape concepts are ignored, and, thirdly, the individual classes often do not overlap with everyday concepts or terms, which are used in natural language descriptions. All three particularities are not to be confused with weaknesses, since the land cover classification still meets its objectives. However, it inhibits from applying this type of landscape description to some potential use cases. One use case where land cover classifications are only of limited applicability is local information retrieval. Land cover classifications do not necessarily correspond with local perception and language use, which are both needed for retrieving information that is of local relevance (e.g. White and Buscher 2012).

We aim at comparing existing land cover classifications with the spatial folksonomy, assuming that the folksonomy uses terms that are frequently used in natural language, and that the spatial folksonomy might have a different spatial focus. We will draw two comparisons. Firstly, we compute the number of different classes used to describe cells of the adaptive grid index (§4.3.3). Secondly, we focus on individual grid cells and qualitatively compare the content of land cover classifications to the descriptions available from the spatial folksonomy. As land cover classifications we incorporate the Arealstatistik (§3.4.1), a fine grained inventory based on sample points organized in a regular grid (100m resolution), and CORINE (§3.4.2), a European initiative, where areas of different land coverage of at minimum 250ha are compiled to a land cover map of scale 1:100,000, covering all of Europe.

5.3 Results and Interpretation

In this study we set out to compute a spatial folksonomy from natural landscape descriptions. The spatial folksonomy consists of a vocabulary of natural features that are frequently used in a Swiss alpine context. The vocabulary is georeferenced and thus reflects local subtleties. In order to achieve this aim we set out

two objectives, as represented in Figure 44, namely annotating a set of prominent natural features, and computing the spatial folksonomy from these features and the adaptive grid index (which resulted from the previous investigation, reported in Chapter 4). The results associated with both tasks will be presented and analyzed in the following sections.

Additionally, we will focus on results from using the spatial folksonomy in order to compare different landscapes and landscape descriptions for similarities. This allows us to answer questions such as *How different is the description of Finsteraarhorn from the description of Uetliberg?* or, more abstractly, *What different types of landscapes can be identified in Switzerland, in terms of their description?* In a last step, we compare the content of the spatial folksonomy to land cover classifications and draw conclusions from answering the question *How can land cover classifications benefit from landscape descriptions?*

5.3.1 Natural Features

From the 1500 most frequent nouns in Text+Berg, 137 were denoted as natural features by at least three out of four annotators and, after linguistic stemming, 94 unique tokens remained (

Appendix B). In Figure 50 the 30 most frequent natural features are graphed and sorted by frequency with English translations. It is important to note that these translations may not be exact matches, but we provide translations to aid understanding.

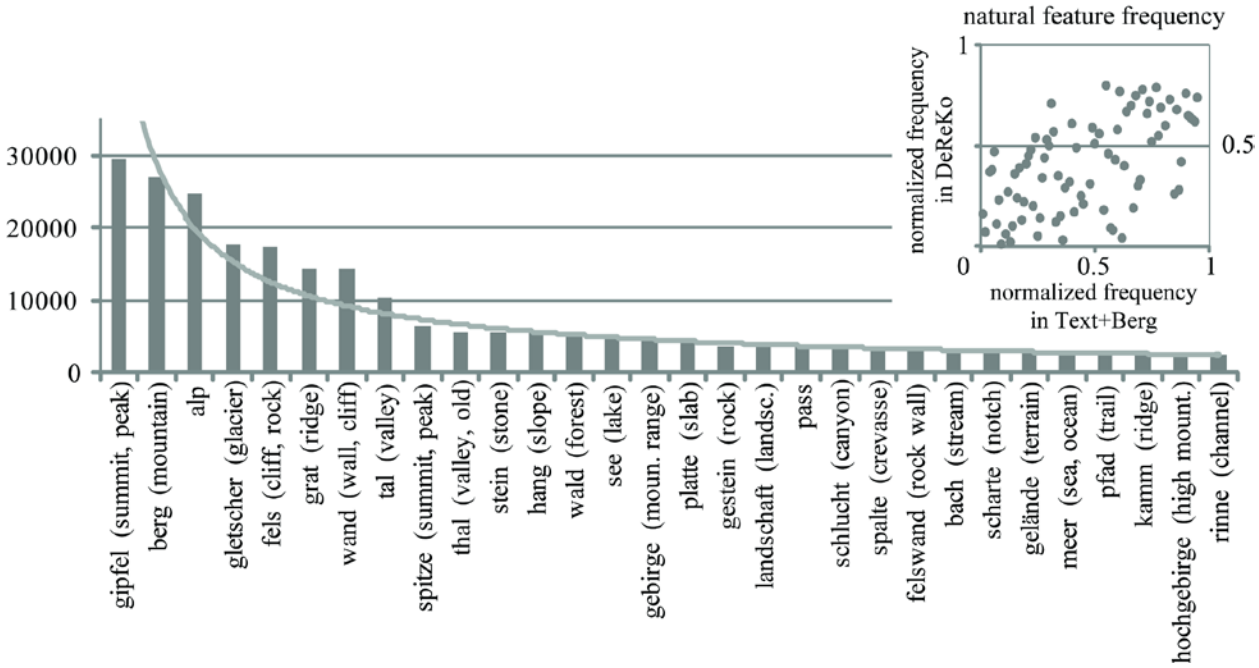


Figure 50. The 30 most frequent natural features in Text+Berg fitted to a quadratic function (r2=0.94). The inset graphs compares frequencies of terms in Text+Berg against frequencies in a general German corpus (DeReKo: §3.2.4).

Natural features in Text+Berg form a detailed vocabulary describing Swiss mountain landscapes. Within the vocabulary we find terms referring to larger regions, such as *Landschaft* or *Gebirge* and terms that are of fine spatial granularity such as *Scharte* (notch), *Spalte* (crevasse) or *Schlucht* (canyon).

The inset in Figure 50 shows that rank of natural features in Text+Berg is not a predictor for rank in a standard German corpus, indicating that Text+Berg's use of German diverges from the norm, and in particular that the terms identified as natural features have some special properties within our corpus. As expected, the frequencies fit well to a Zipfian distribution (r^2=0.94, 50000*x^{-1}, Figure 50), with almost all natural features being related to Swiss mountains – of the top 30, only *Meer* (sea or ocean) is an exception. Since some of the 94 natural features that do not match the mountain context of the Text+Berg corpus, and since such features become slightly more frequent towards the end of the natural feature list (

Appendix B), we thus argue that by considering more than the 1500 most frequent nouns we might not gain much more information, since the number of mountain irrelevant features would also increase.

In Figure 51 we analyze the distribution of annotated natural features over all descriptions in Text+Berg, by comparing the frequency ranking of natural features with a frequency ranking, reflecting the number of descriptions that contain the respective feature (normalized ranks, rank 1 = most frequent).

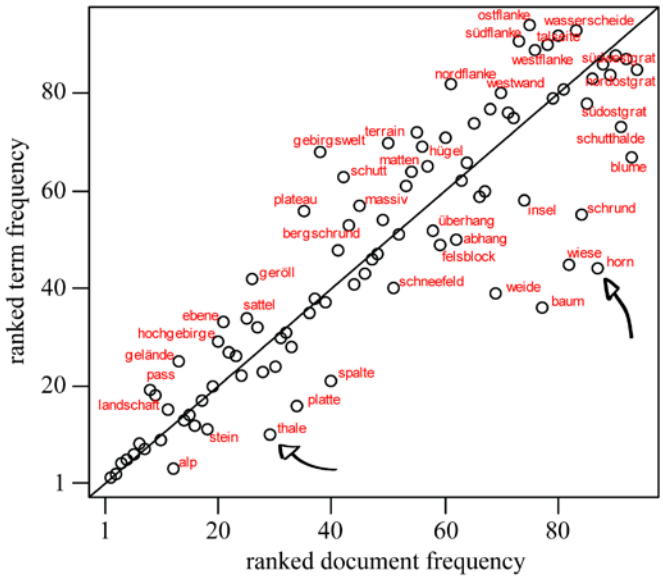


Figure 51. Comparison of frequency of natural features in the corpus and their distribution over all documents (below diagonal line = distributed over only few documents)

The overall correlation between the two frequency rankings in Figure 51 is 0.88 (Spearman rho), indicating that natural features are largely equally distributed over all documents. However, there are some exceptions to this rule, such as *thale*, *platte*, *spalte*, *baum*, *horn*, *wiese*, *weide* and *schrund*, all located below the cross section drawn in Figure 51. These natural features are particular for only a small subset of documents and thus not representative for the whole corpus. Most of these exceptions are features that refer to landscapes not particular for a *Swiss* mountain context, such as *Wiese*, *Weide* and *Matten*, all referring to agricultural fields. Another group of unequally distributed features change over time. Two examples are *Thal* (valley) and *Horn* (peak, summit). *Thal* is the old spelling of *Tal* (valley) and thus only used in early descriptions[39]. *Horn* was used in early yearbooks for referring to the summit of mountains and sometimes to mountains as wholes. However, *Horn* does not occur in more recent

[39] We did not merge the two spellings *thal* and *tal*. One reason is that we initially wanted to have a closer look at changes of descriptions over time. For this reason it would have been interesting to see if the two spellings of tal are associated with different concepts. Additionally, *thal* and *tal* could be considered synonyms. We did not aggregate any of the other synonyms in the natural feature list. We only aggregated lexemes.

documents in this same context. We analyzed and discussed the *Horn* example in a detailed study, published in Derungs et al. (2013).

Previous empirical work has concentrated on identifying basic levels or category norms by conducting empirical investigations where participants were asked to list *natural earth formations* (Battig and Montague 1969, Van Overschelde 2004) and *a kind of geographical feature* (Smith and Mark 2001). Table 4 shows the ranking of the natural features identified in Text+Berg in comparison with the top terms identified in these experiments. The last column shows terms that were prominently used to describe photographs uploaded to Geograph[40] (Edwardes *et al.* 2007).

Table 4. Top 20 basic levels and category norms from different investigations and their respective frequency rank, if existing, from Text+Berg.

1		**2**		**3**		**4**	
Battig & Montague	**Rank (T+B)**	**Van Overschelde et al.**	**Rank (T+B)**	**Smith & Mark**	**Rank (T+B)**	**Edwardes & Purves**	**Rank (T+B)**
Mountain	2	Mountain	2	Mountain	2	Road	NA
Hill	69	River	41	River	41	Hill	69
Valley	8	Ocean	26	Lake	14	River	41
River	41	Volcano	NA	Ocean	26	Village	NA
Rock	5	Lake	14	Hill	69	Building	NA
Lake	14	Valley	8	Plain	33	Park	NA
Canyon	20	Hill	69	Plateau	56	Street	NA
Cliff	7	Rock	5	Desert	NA	Valley	8
Ocean	26	Canyon	20	Volcano	NA	Field	62
Cave	NA	Plateau	56	Island	58	Loch	14
		Tree	36	Plain	33	Land	NA
		Plain	33	Plateau	56	Town	NA
		Cave	NA	Map	NA	Forest	13
		Glacier	4	Road	NA	Map	NA
		Grand Canyon	NA	Island	58	Sea	NA
		Island	58	Desert	87	Woodland	13
		Stream	23	Peninsula	NA	Tree	36
		Cliff	5	State	NA	Beach	NA
		Desert	87	Volcano	NA	Country	NA
		Beach	NA	Forest	13	Glen	NA

Three key points can be made when comparing the annotated natural features with findings from empirical investigations on relevant geographic objects (column 1, 2 and 3). Firstly, most of the identified terms in previous empirical work are also represented in Text+Berg, with a few exceptions such as *beach*, *cave*, *volcano* and *desert*. Most of these features are usually not associated with a *Swiss* mountain context,

[40] www.geograph.org.uk

while it appears that caves are rarely mentioned in the corpus. Secondly, the frequencies of terms used in Text+Berg have little relationship with those suggested by the participants in previous experiments, reflected by unsorted rankings of Text+Berg columns. Thirdly, many of the most frequent natural features in Text+Berg were not listed in the top 10 categories in previous work; examples include *summit*, *alp*, *glacier* and *arête*. Some of these might be considered new basic levels in a (Swiss) alpine context, e.g. *glacier* or *alp*. Other prominent features might represent sub- or super-ordinates of known basic levels, such as *mountain range* (super-ordinate) or *rock wall* (sub-ordinate). A rather large set of natural features appears to match with known basic levels. *Summit*, *arête* and *ridge* could all be considered *proper parts* of the feature *mountain*. The classification of features into basic levels, sub- or super-ordinates and parts or wholes is a challenging task and the data we extracted from Text+Berg might not be sufficiently rich to allow such structuring. However, the annotated natural features from Text+Berg appear to give us access to an extensive and detailed vocabulary describing (Swiss) mountain landscapes at a fine level of spatial granularity.

The comparison with prominent terms used in Geograph shows a slightly different picture. Most frequently used terms in Geograph are not available as natural features in Text+Berg. This is mainly due to the prominence of artificial features in Geography, for example *road*, *village* or *building*. Artificial features were explicitly ignored in the annotation process for retrieving natural features from Text+Berg, where one annotation rule requires natural features to be natural (§5.2.1). However, there is some overlap between natural features and frequent terms in Geograph, for example *hill*, *valley* or *forest*. All of these examples are also resolved in the empirical investigations and thus considered basic level categories.

An obvious limitation of the annotation process is the limited control over semantic ambiguity. Some terms identified as natural features might be much more commonly used in a non-natural feature context. One example is *wand* (wall), which refers to a mountain face (as in *big wal*") as well as to a mundane wall of a building. The annotators were informed on the context of the corpus and thus, a majority of the annotators identified *wand* as a natural feature.

We encountered two problems when comparing natural features with prominent geographic objects, stemming from empirical investigations. Firstly, the translation from German nouns in Text+Berg to English terms as published in the empirical investigations is critical. Some examples are very unclear, as for instance *woodland* and *forest*, which are both *wald*. Sometimes whole groups of natural features are critical, as for instance water streams. The English language has comparably more terms that refer to water streams (e.g. *fork*, *kill*, *lick*, *stream* or *gill*), whereas German has only a few terms available. In many cases there is no one-to-one relationship between different terms in different languages, which

makes comparisons difficult. This is not too surprising, and is also reflected in ethnophysiographic literature, where local variation of landscape terms has been shown to be significant.

Secondly, the nature of the data gathered through empirical investigations is clearly different from the data retrieved from natural language descriptions. By comparing the different lists of terms we assume that term frequency in a corpus of natural landscape descriptions is comparable to the ranking of terms when for instance asked for examples of *geographic concepts*. The two sources for landscape terms might be roughly comparable. However, in many cases the context of a description must be considered significantly different from the situation given in an empirical experiment. Thus, the discussion of similarities and differences between natural features and results from empirical investigations is to be considered with caution.

5.3.2 Spatial Folksonomy

In the following section we will review the characteristics of the spatial folksonomy and its suitability for drawing qualitative and quantitative comparisons between regions. As a means of detailed qualitative comparisons, we will have a closer look at the content of distinct grid cells. In order to perform large scale quantitative comparisons, we compute similarities between cell vectors, as described in Figure 48, and thus generate similarity surfaces for the extent of whole Switzerland.

5.3.2.1 Qualitative Comparison of Landscape Information Retrieved for Different Regions

We investigate the content, i.e. the frequency distribution of the 5 most frequent and most particular natural features for a set of 12 distinct grid cells, describing different regions. Some of these regions contain natural features as diverse as mountains (e.g. *Matterhorn* or *Finsteraarhorn*), villages (e.g. *Thun* or *Lenzerheide*) and valleys (*Toggenburg*). Additionally, the 12 cells have different grid size and they are distributed all over Switzerland. The label of each region is added manually and reflects what we consider a suitable description of its content. Often the label reflects the name of a mountain or a valley. In Figure 52 the 5 most frequent and particular natural features, with respect to feature counts (*tf*) and *tf-idf* values (Equation 1), are listed for all 12 regions.

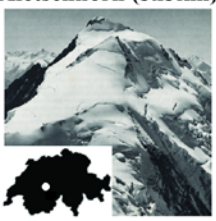
Matterhorn (5x5km)

tf	*tf-idf*
berg	berg
gipfel	grat
grat	tal
gletscher	stein
tal	spitze

Finsteraarhorn (5x5km)

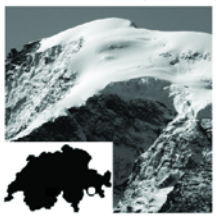
tf	*tf-idf*
gipfel	gletscher
gletscher	spitze
berg	grat
grat	horn
spitze	vorgipfel

Aletschhorn (5x5km)



tf	*tf-idf*
gipfel	gletscher
gletscher	platte
grat	stein
berg	grat
wand	baum

Morteratsch (5x5km)



tf	*tf-idf*
gipfel	spitze
gletscher	eiswand
berg	grat
spitze	gletscher
grat	scharte

Tödi (5x5km)

tf	*tf-idf*
berg	gestein
gipfel	gletscher
grat	alp
gletscher	grat
tal	felswand

Salbit (5x5km)

tf	*tf-idf*
gipfel	platte
grat	westgrat
gletscher	grat
wand	wand
berg	südgrat

Cristallina (5x5km)

tf	*tf-idf*
alp	alp
gipfel	tal
tal	grat
grat	bach
berg	pfad

Lenzerheide (10x10km)

tf	*tf-idf*
alp	alp
berg	tal
tal	berg
gipfel	pass
pass	spitze

Uetliberg (5x5km)

tf	*tf-idf*
alp	baum
berg	fluss
tal	wald
wald	tal
see	see

Thun (20x20km)

tf	*tf-idf*
gletscher	gletscher
alp	einzugsgebiet
gipfel	see
berg	gletscherzunge
grat	alp

Toggenburg (10x10km)

tf	*tf-idf*
alp	grat
gipfel	alp
grat	wand
berg	schlucht
wand	gestein

Wendenstöcke (10x10km)

tf	*tf-idf*
gipfel	wand
alp	platte
berg	eiswand
wand	alp
gletscher	see

Figure 52. Top 5 natural features, with respect to feature count (*tf*) and tf-idf values, for 12 different regions.

Feature counts (*tf*) are fairly similar for most of the 12 regions (an issue that we earlier discussed as a consequence of Zipf distributed values, §5.2.2). The feature *berg* (mountain) occurs in all 12 regions,

gipfel (summit) in 11 and *gletscher* (glacier) in 8 out of 12 examples. Even in *Thun*, a city located at a lake, adjacent to the *Alps*, we count many occurrences of *gletscher*, *berg* and *gipfel*. These features are not only prominently represented in the 12 regions shown in Figure 52 but also occur in the top 10 list gathered from the whole corpus, as shown in Figure 50. Thus, these features cannot be considered specific local descriptors and we will consequently not use feature counts in further investigations on particularities of regions.

More specific descriptions emerge from considering relative feature counts, i.e. *tf-idf values*. The comparison of tf-idf values associated with the 12 regions, as represented in Figure 52, indicate that the spatial folksonomy can describe local landscapes on a level of great detail. The region of *Matterhorn*, for instance, is clearly dominated by the mountain itself. Thus, *Matterhorn* is described by mountain related features such as *berg* (mountain), *grat* (ridge) and *spitze* (summit), but also *tal* (valley) reflecting its abrupt emergence at the end of *Mattertal*. *Finsteraarhorn*, on the other hand, is a prominent mountain embedded in the glacier landscape of the *Bernese Oberland*. In terms of natural features, *Finsteraarhorn* is described by *gletscher* (glacier) and a set of peak related features, such as *spitze* and *vorgipfel*. *Salbit* is known for its quality granite and related rock-climbing, with two famous ridges emerging as important terms (*west* and *südgrat*), as well as the generic term *grat* (ridge). *Thun* and *Uetliberg* are both characterized by terms which might be more commonly related to lower, more accessible regions, examples are *see* (lake), *alp* (alps), *wald* (forest) and *baum* (tree).

5.3.2.2 Quantitative Comparison of Landscape Information Retrieved for Different Regions

In a first quantitative comparison of the landscape information stored in the spatial folksonomy we compare the frequency distribution of all 94 natural features of the 12 regions discussed above. As a measure of similarity we compute cosine similarities in parallel with statistical dependence.

Table 5 and Table 6 show cosine similarities calculated for all pair-wise comparisons of the 12 regions. Cosine similarities are calculated separately for feature counts (

Table 5) and the tf-idf values (Table 6) associated with all 94 natural features (not only the top 5 features, as qualitatively inquired above).

Table 5. Cosine similarities between the natural feature term frequencies of 12 different regions.

	Matterhorn	Finsteraarhorn	Aletschhorn	Morteratsch	Tödi	Salbit	Cristallina	Lenzerheide	Uetliberg	Thun	Toggenburg	Wendenstöcke
Matterhorn	1											
Finsteraarhorn	.9	1										
Aletschhorn	.8	1	1									
Morteratsch	.9	.9	.9	1								
Tödi	.9	.9	.9	.9	1							
Salbit	.8	.9	.9	.9	.9	1						
Cristallina	.9	.8	.7	.8	.8	.8	1					
Lenzerheide	.9	.7	.7	.7	.7	.7	.9	1				
Uetliberg	.7	.4	.4	.5	.5	.4	.7	.8	1			
Thun	.7	.8	.8	.7	.8	.7	.6	.6	.5	1		
Toggenburg	.9	.9	.9	.9	.9	.9	.9	.8	.6	.7	1	
Wendenstöcke	.9	.9	.9	.9	.9	.9	.8	.8	.6	.7	.9	1

Table 6. Cosine similarities between the tf-idf values of 12 different regions. Grey shaded tf-idf values are statistically independent.

	Matterhorn	Finsteraarhorn	Aletschhorn	Morteratsch	Tödi	Salbit	Cristallina	Lenzerheide	Uetliberg	Thun	Toggenburg	Wendenstöcke
Matterhorn	1											
Finsteraarhorn	.6	1										
Aletschhorn	.5	.7	1									
Morteratsch	.6	.6	.6	1								
Tödi	.6	.7	.7	.6	1							
Salbit	.5	.5	.6	.5	.6	1						
Cristallina	.5	.3	.4	.3	.5	.4	1					
Lenzerheide	.6	.4	.3	.3	.4	.2	.6	1				
Uetliberg	.3	0	.1	0	0	0	.3	.4	1			
Thun	.3	.4	.4	.3	.5	.3	.3	.4	.3	1		
Toggenburg	.6	.4	.5	.5	.7	.6	.7	.5	.3	.5	1	
Wendenstöcke	.6	.4	.5	.5	.6	.5	.4	.3	.2	.4	.6	1

Table 5 and Table 6 indicate that the spatial folksonomy can be used to compute quantitative similarities between regions, as it's described in §5.2.3. The pair-wise cosine similarities between the 12 regions are

in accordance with the conclusions we draw from qualitatively comparing the top 5 natural features, as listed in Figure 52. Firstly, similarities between counts (tf, Table 5) are generally higher compared to similarities between tf-idf values (Table 6). Similarities between tf-idf values have a wider spectrum, compared to similarities between feature counts, and thus better represent particular relations and differences between regions. Secondly, the cosine similarities, in particular between tf-idf values (Table 6), meet our expectations, for instance gained from comparing the photographs of the 12 regions. *Matterhorn* is for instance similar to *Finsteraarhorn* (0.6) and different from *Uetliberg* (0.3).

Manual comparisons between the top 5 natural features of all regions, as discussed above, seem to be well suited in order to *understand* differences between the 12 regions, whereas numeric comparisons show less explicit results, for instance reflected by significant correlations, given for almost all examples (only the grey shaded similarities in Table 6 are uncorrelated). Thus, *Matterhorn*, *Finsteraarhorn* and *Aletschhorn* are statistically related to all other regions, with only one exception, namely *Uetliberg*. This is clearly surprising, since many of the other 8 regions, besides *Uetliberg*, are represented by quite different landscape characteristics, compared to the three prominent mountains. Additionally, some of the quantitative comparisons are counter intuitive. One example is given by comparing the *Lenzerheide-Matterhorn* similarity with the similarity between *Piz Morteratsch* and *Salbit*. Both pair-wise comparisons show the same similarity value, namely 0.6. Thus the similarity between a mountain village (*Lenzerheide*) and the most prominent mountain in *Switzerland* (*Matterhorn*) is supposedly equal to the similarity between two mountains (*Piz Morteratsch* and *Salbit*). In summary, the quantitative comparisons seem to generate meaningful results on a broad scale. However, individual comparisons can be unexpected and sometimes wrong.

5.3.2.3 Spatially continuous landscape similarity

In Figure 53 we show means for answering the question *How different is the description of Uetliberg from Finsteraarhorn?*, and thus compare the description of the two regions, using cell vectors consisting of tf-idf values of all 94 natural features. We thus compute *landscape similarity maps* for both regions, *Uetliberg* and *Finsteraarhorn* (§5.2.3). Importantly, documents that contribute terms to the target cells are not used in all other cells, which allow control of spatial autocorrelation.

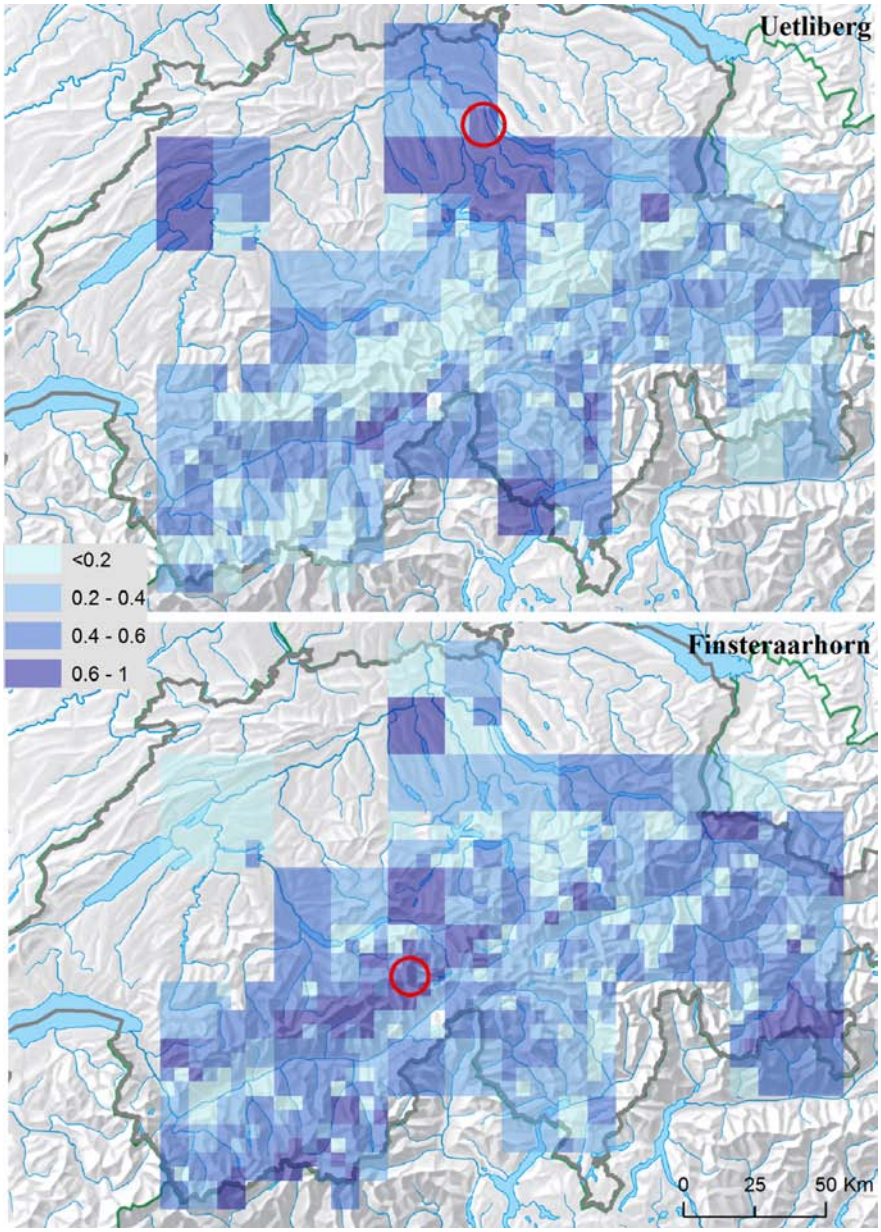


Figure 53. Landscape similarity maps for Uetliberg and Finsteraarhorn (red circles), computed from cosine similarities between tf-idf values of all natural features and for cells of the spatial folksonomy.

The patterns shown by the two landscape similarity maps meet our expectations. The similarities associated with the two regions show an inverse pattern, with for instance the *Bernese Oberland* and the *Valais Alps* having similar descriptions compared to *Finsteraarhorn* (bottom map, dark blue), and almost no similarity with the description of *Uetliberg* (top map, bright blue). *Uetliberg*, on the other hand, shows high similarity with regions at the foothill of the *Alps* and with broad valley floors. The inverse trend between the two maps is also reflected by a negative correlation of -0.32 (Spearman rho). The similarity computations are generally more representative for regions for which the Text+Berg corpus provides rich

descriptions. Peripheral regions, such as parts of the Swiss *Mittelland* or the *Jura*, are often noisy, in terms of showing unpredictable similarity values.

5.3.2.4 Explaining the Variation in Landscape Information

The computation of landscape similarity maps introduced additional means for interpretation, compared to the qualitative comparisons, as shown in Figure 52. However, the answers to questions on how these maps could be evaluated or how differences between landscape information can be explained cannot be deduced from the landscape similarity maps only. For this reason we compare the variation of landscape information of the two regions *Uetliberg* and *Finsteraarhorn* with the variation of an explanatory variable, namely the variation of geomorphometric characteristics. In contrast to descriptions, which we first have to georeference and resolve from text, geomorphometric characteristics can be considered robust. Geomorphometric characteristics are independent from perception and only represent the shape of the earth's surface. Similar shape will always be expressed by similar geomorphometric characteristics. Geomorphometric similarities are used to compute geomorphometric similarity maps (§5.2.3.1). These maps are then qualitatively and quantitatively compared to the landscape similarity maps as shown in Figure 53. Figure 54 shows landscape and geomorphometric similarity maps for Uetliberg and Finsteraarhorn (the landscape similarity maps are reused from Figure 53).

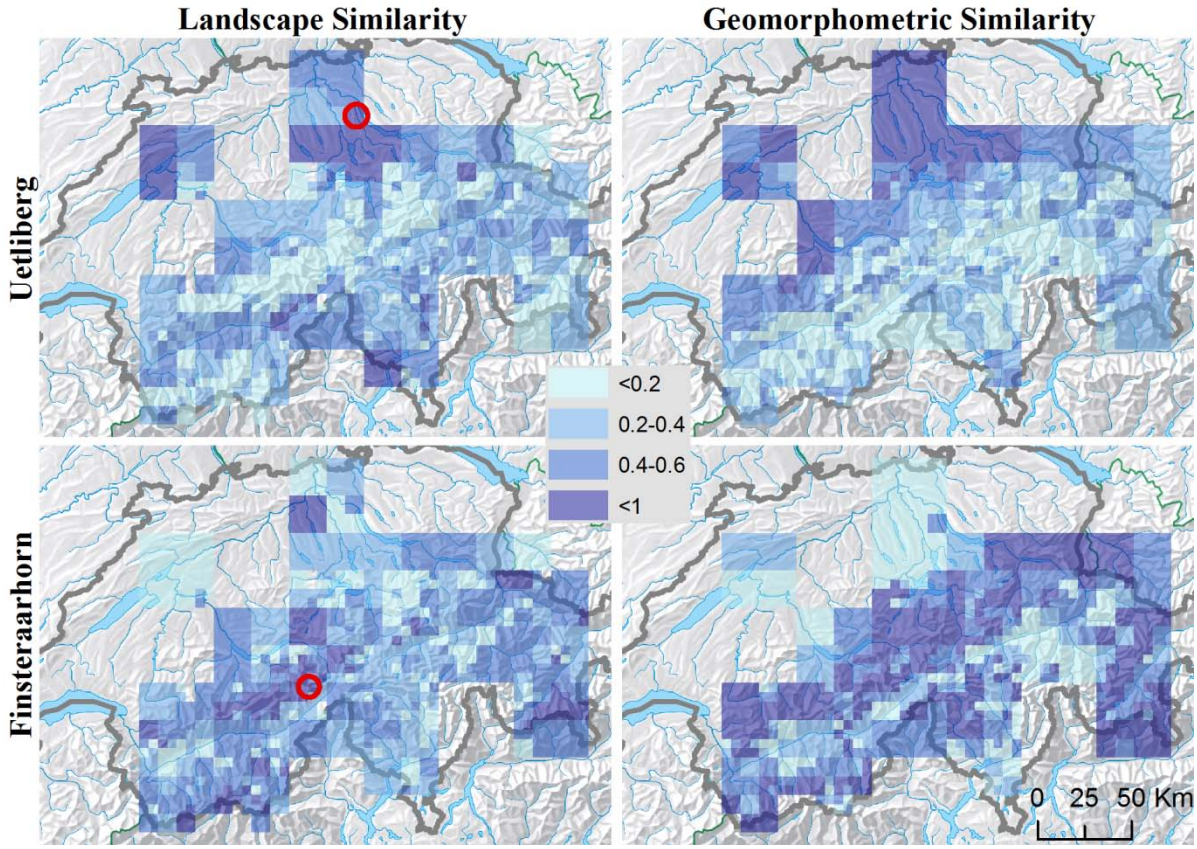


Figure 54. Landscape and geomorphometric similarity maps for Uetliberg and Finsteraarhorn (red circles).

The two types of similarity maps, based on similar descriptions (Figure 54, left) and similar geomorphometric characteristics (right), show related patterns. This indicates that the variation of topography is an expressive descriptor of the variation found in descriptions. Or, the descriptions in Text+Berg seem to be influenced by surrounding geomorphometric characteristics. This is clearly reflected by correlation values between the two types of maps as shown in Table 7.

Table 7. Correlation (Spearman rho) of the landscape (LAND) and geomorphometric (GEOM) similarity maps of Uetliberg and Finsteraarhorn.

	***LAND* Uetliberg**	***LAND* Finsteraarhorn**
***GEOM* Uetliberg**	0.27	-0.04
***GEOM* Finsteraarhorn**	-0.34	0.43

The landscape and geomorphometric similarity maps of *Finsteraarhorn* correlate with a coefficient of 0.43 (Spearman rho). The similarity maps of *Uetliberg* correlate with 0.27 (Spearman rho). Both correlations are statistically significant, meaning that topographic similarities cannot be considered independent from similarities between descriptions. The higher correlation between the description and

topography of *Finsteraarhorn*, compared to *Uetliberg*, indicates again that the corpus contains more reliable descriptions of landscapes in the Swiss Alps.

For means of comparison we correlated the geomorphometric similarity of *Finsteraarhorn* with the landscape similarity of *Uetliberg* and vice versa. From these comparisons we gained negative or very low correlations, namely -0.34 and -0.04. This reflects that similarities with the description of *Finsteraarhorn* are inversely related to the similarities with the topography of *Uetliberg*. The description of *Uetliberg* is unrelated to the topography *Finsteraarhorn*. All four correlations nicely reflect that the comparison between landscape descriptions and geomorphometric characteristics can be considered as a means for explaining the variation in how people perceive and describe natural mountain landscapes in Switzerland.

The correlation between descriptions and geomorphometric characteristics is interesting for three reasons. Firstly, as mentioned above, the correlation suggests new means for explaining variation in descriptions, namely by the shape of the earth's surface. Secondly, the correlation is surprising since there is no explicit link between georeferenced and structured text documents and digital elevation models. The two data sets are entirely independent but still show significant correlation. This leads to the third point of interest, namely potential means of further investigations. By correlating semantically rich descriptions with geometrically rich terrain data we could combine the best of the two worlds in one predictive model, which could potentially allow for automatically deduced meaningful local descriptions from shape.

5.3.2.5 From Landscape Information to Landscape Typology

Instead of computing pair-wise similarities between individual cells of the spatial folksonomy, we can also apply the whole spatial folksonomy to a clustering and thus automatically create groups of similarly described regions. The associated question with this approach would be: *What different types of landscapes can be identified in Switzerland, in terms of their description?* In Figure 55 we clustered the spatial folksonomy, i.e. all cell vectors, into 2, 4 and 8 landscape groups.

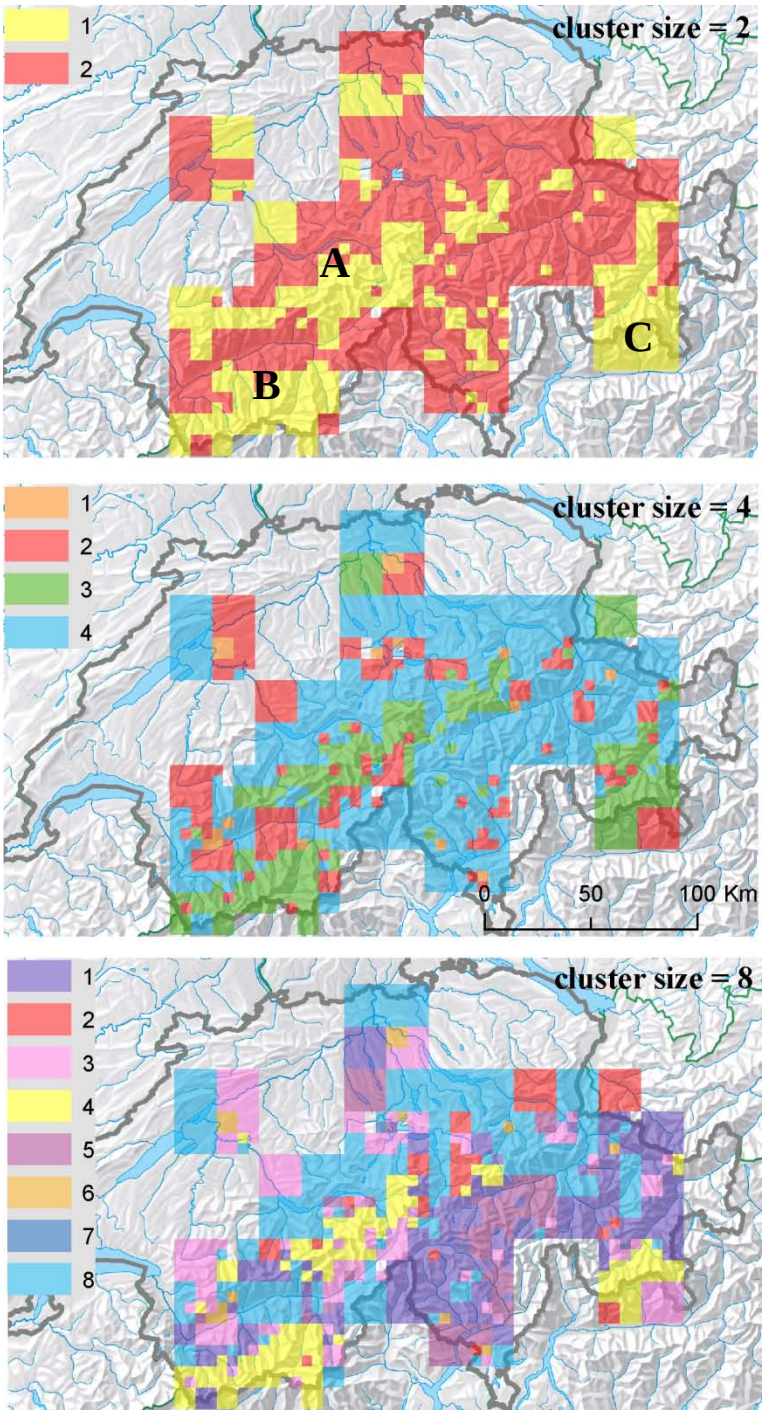


Figure 55. K-means clustering of all cell vectors (<40km resolution) for three cluster sizes (2, 4 and 8).

The maps in Figure 55, generated from clustering all cell vectors of the spatial folksonomy, highlight regions of similar description, by using a similar color code. Some regions, such as the *Bernese Oberland* (A), the *Valais Alps* (B), and the *Bernina* (C) region are consistently grouped together, independent of cluster size and Euclidean distance. Additionally, focusing on the map created from using k=4, we have

the means for illustrating how the spatial folksonomy can be used to automatically group landscapes in Switzerland into meaningful entities, such as high alpine regions (green), regions that border with these (red) and the rest of the Swiss Alps (blue).

In order to test the hypothesis that clustering applied to the spatial folksonomy results in meaningful landscape groups, we intersect the results with an official typology of Swiss landscapes (§3.4.3). The typology of Swiss landscapes distinguishes between five types of landscapes; the Swiss *Mittelland*, zones that are particular warm or low (*tief oder warm*), the *Jura*, pre-alpine regions (*Vorgebirge*) and high alpine regions (*Hochgebirge*) (§3.4.3 and Figure 22). Figure 56 is an overlay of the map produced by clustering landscape descriptions of Text+Berg into four groups and the above introduced Swiss landscape typology.

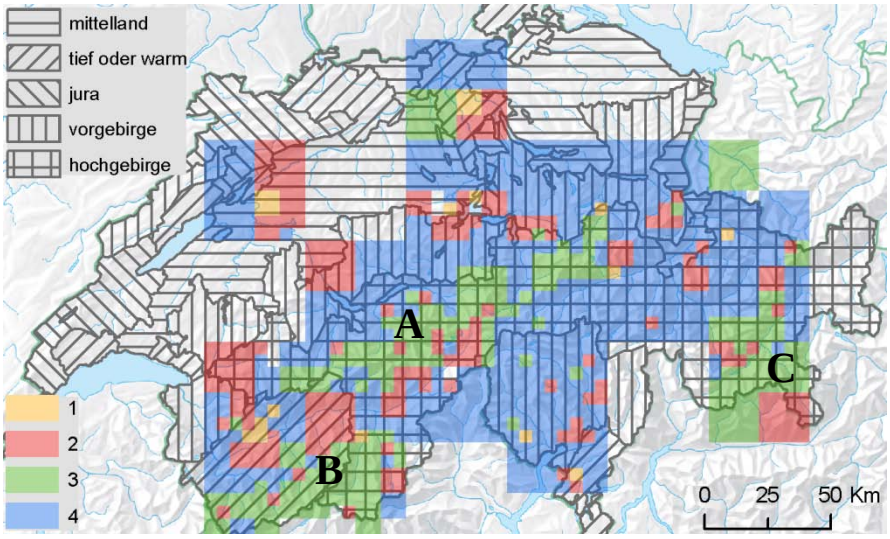


Figure 56. Comparison of landscape types generated through clustering (color schema, *k*=4) and provided by an official landscape typology (background pattern, §3.4.3).

The two maps seem to be unrelated. It appears that the green cluster (3), which covers the *Bernese Oberland* (A), the *Valais Alps* (B) and the *Bernina* (C) region, shows most overlap with the landscape type *Hochgebirge*. This is reasonable. However, the blue cluster (4) also overlaps with *Hochgebirge*, as well as with the landscape type *Vorgebirge*. This quite complex relationship between individual clusters and landscape types is also reflected in Figure 57, where the relative distribution of each cluster - with cluster sizes 2, 4 and 8 - over all landscape types is summarized.

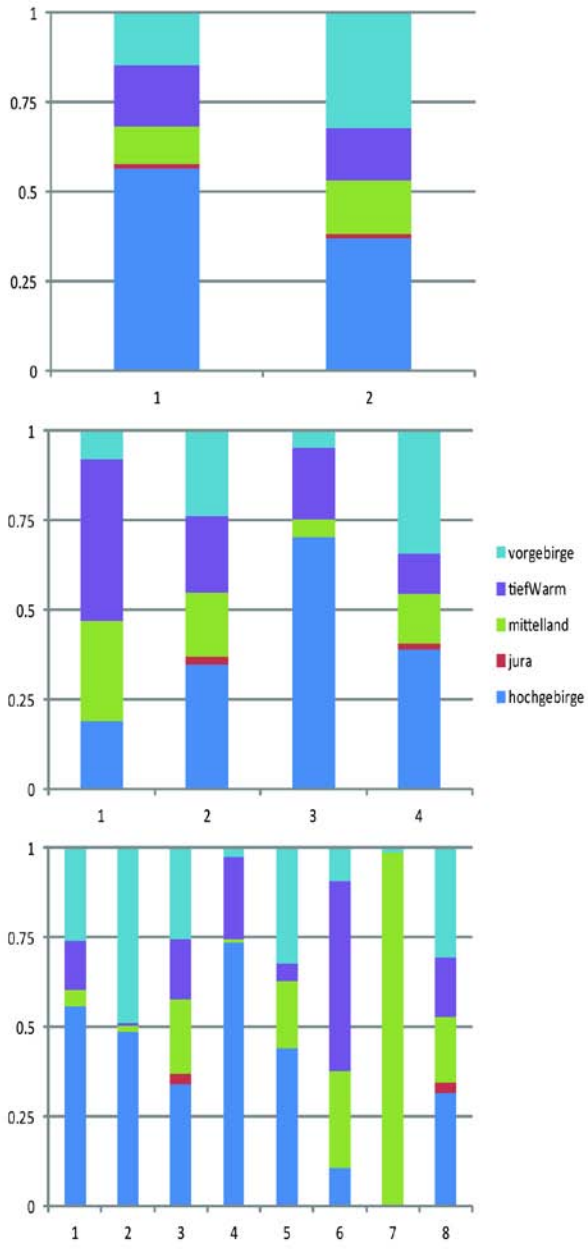


Figure 57. Relative distribution of clusters on the five types of Swiss landscapes.

The *Hochgebirge* landscape type covers large areas of the Swiss Alps, and is the region best described in the spatial folksonomy. Thus, applying clustering to the spatial folksonomy mainly leads to a segregation of the *Hochgebirge* landscape type into different subregions. This is clearly visible in Figure 57, where *Hochgebirge* is the dominant landscape type for most clusters, independent of cluster size. A vivid example is given by clustering of the spatial folksonomy into two landscape groups, which are both dominated by the landscape type *Hochgebirge*. Thus, the official Swiss landscape typology has too coarse resolution to be comparable to the spatial folksonomy.

The question of whether clustering the spatial folksonomy results in meaningful landscapes remains without a distinct answer. A visual interpretation of the different clusters, and that many regions remain consistent over different cluster sizes, suggests that clustering indeed is a means for generating meaningful landscapes. However, the intersection with the Swiss landscape typology lead to the conclusion that the official Swiss landscape typology needs better resolution for Swiss alpine regions in order to provide the means for comparisons with the spatial folksonomy.

The different resolutions of the two types of landscape typologies, one officially used in Switzerland and mainly deduced from land cover classifications and the other computed from the spatial folksonomy, which consists of information from descriptions of mountain landscapes as perceived by people, bears an important insight. Describing landscapes, for instance by describing outdoor activities, leads to more detailed and diversified landscape typologies, compared to the one that is available for the whole of Switzerland and officially used in political decision making processes.

5.3.3 Folksonomy and Land Cover Classifications

The question that guided the comparison between land cover classifications and the spatial folksonomy was whether the two types of landscape descriptions could profit from one another. We thus compared the spatial focus and the semantic characteristics of the descriptions separately.

The first comparison was on the spatial focus of two land cover classifications, namely Arealstatistik (§3.4.1) and CORINE (§3.4.2), compared to the spatial folksonomy. We thus compared the (relative) number of classes that are applied to classify the content of cells of the adaptive grid (Figure 58). The colors refer to the numbers of classes used (< 20% means that less than 20% of all available classes are used for this particular cell, which for Arealstatistik means an equivalent of 15 classes (total n classes = 72), 8 classes for CORINE (n = 44) and 18 classes for the spatial folksonomy (n = 94)).

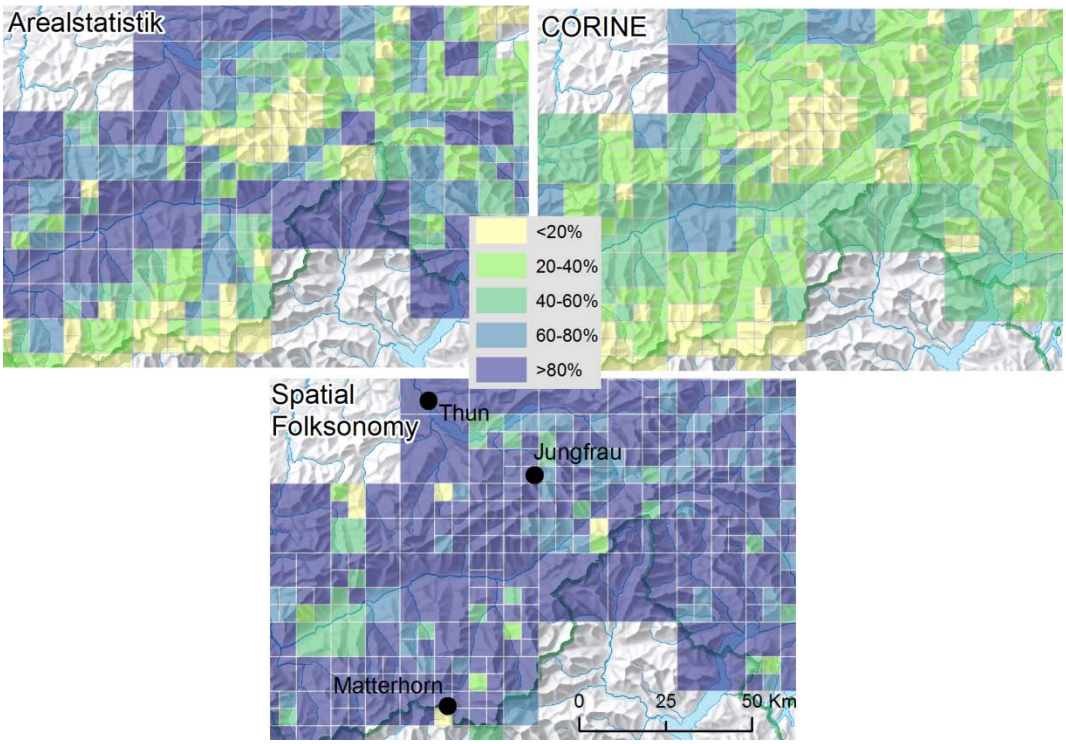


Figure 58. Classification diversity of two land cover classifications, Arealstatistik (upper left) and CORINE (upper right), and the spatial folksonomy (bottom), in terms of relative numbers of classes available for cells of the adaptive grid.

Unsurprisingly, the resolution of grid cells has impact on the number of classes. However, it is also obvious from Figure 58 that CORINE and the Arealstatistik both have their spatial focus on populated places and settlement areas, indicated by high numbers of classes used to classify valley floors and only a few classes being available to label high alpine regions such as the *Bernese Oberland* (e.g. Jungfrau) or the *Valais Alps* (e.g. Matterhorn). This is contrasted by the pattern evolving from the spatial folksonomy, where almost all natural features are used to describe core regions in the *Alps* and fewer for regions in the valley bottom or at the foothill of the *Alps*. The spatial folksonomy and the two land cover classifications are complementary in terms of spatial coverage.

In Figure 59 we compare relative numbers of classes applied to the previously discussed 12 distinct regions (c.f. Figure 52).

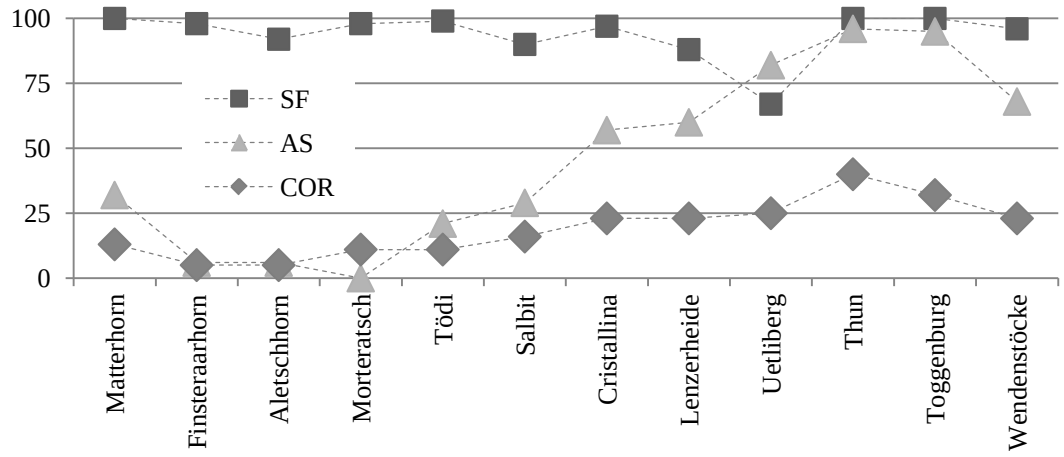


Figure 59. Relative numbers of classes available in the spatial folksonomy (*SF*), Arealstatistik (*AS*) and CORINE (*COR*) to describe 12 regions.

The spatial folksonomy (*SF*) makes use of almost 100% of all available natural features for many cells of the adaptive grid. In Figure 59 we see that in 12 topographically diverse regions, only *Uetliberg* and *Lenzerheide* are described by using less than 90% of the 94 available natural features. The use of almost 100% of the available vocabulary for most cells is due to the nature of the spatial folksonomy, where information on the earth's surface is retrieved from counting occurrences of natural features in georeferenced text documents. The probability that a natural feature occurs in one of the text documents associated with a grid cell is relatively high and does not necessarily imply that the respective features *really occurs* at this location. Occurrence is to be considered in combination with term frequencies, in order to deduce meaningful descriptions, as we have shown in many examples in §5.3.2. Land cover classifications, on the other hand, directly link to the earth's surface, such that the occurrence of a land cover class in a certain region means real occurrence. Figure 59 also relates to the spatial focus of land cover classifications, which is clearly biased towards populated areas, such as villages (*Lenzerheide*), towns (*Thun*) or the *Swiss Mittelland* (*Uetliberg*). Alpine regions are peripheral in the land cover classifications and only described in limited detail and by using only a small subset of the available vocabulary.

In Figure 60 we compare semantic characteristics of the spatial folksonomy, compared to land cover classifications. We thus compare the top 5 tf-idf classes of all three landscape descriptions, applied to the 12 regions, already discussed in Figure 52.

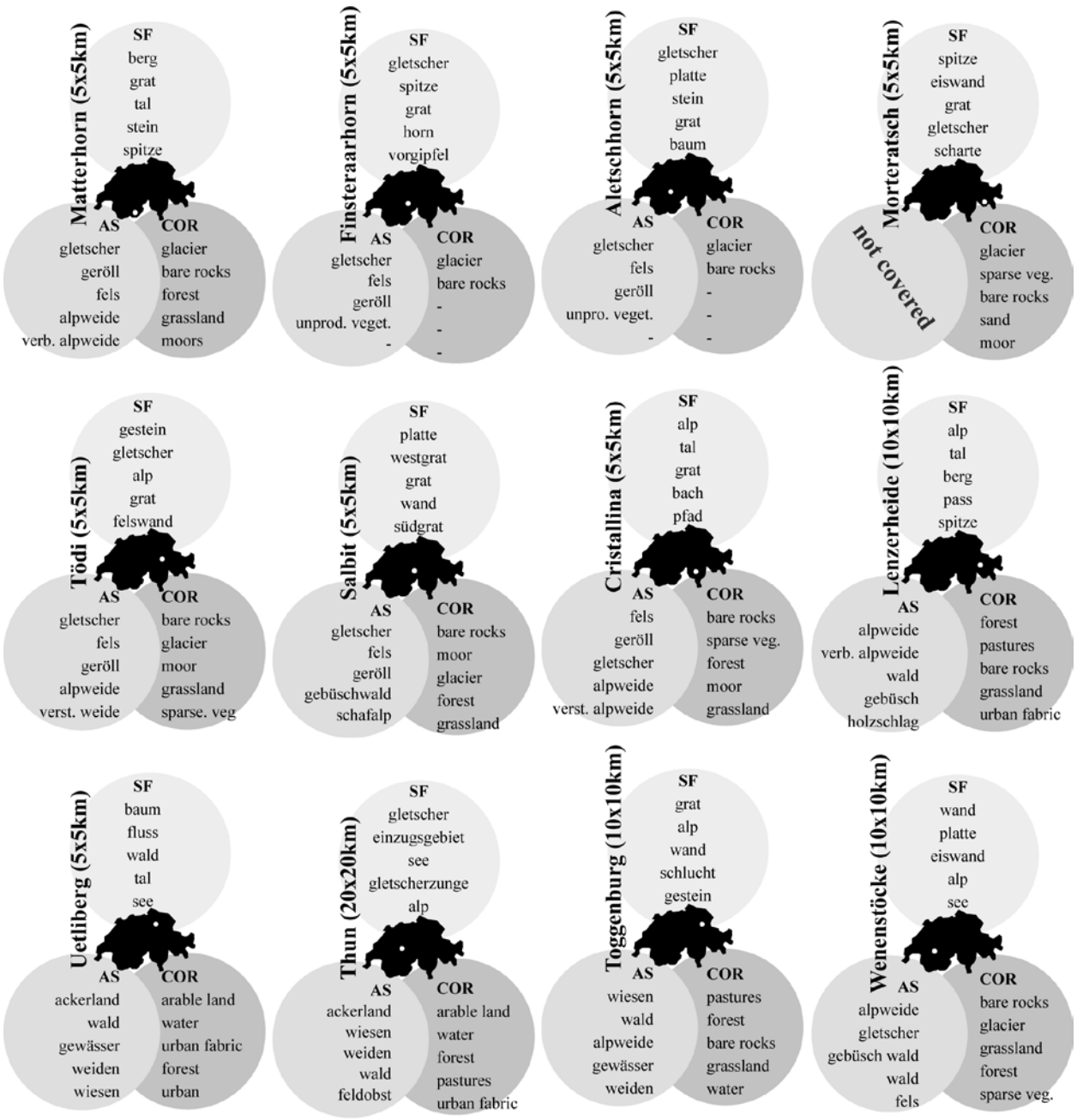


Figure 60. Top 5 spatial folksonomy (*SF*), Arealstatistik (*AS*) and CORINE (*COR*) terms according to tf-idf values, for 12 regions.

Figure 60 indicates that both land cover classifications, the Arealstatistik and CORINE, contain rich and detailed descriptions for regions, such as *Uetliberg*, *Toggenburg* or *Lenzerheide*, referring to different types of agricultural fields (e.g. *weiden*, *wiesen*, *ackerland* or *feldobst*). The availability of classes for describing alpine landscapes, on the other hand, is considerably sparse. The regions of *Aletschhorn* and *Finsteraarhorn*, for instance, are described by the use of only two (CORINE: *glacier*, *bare rock*s) or four

classes (Arealstatistik: *gletscher, fels*, *geröll*, *unproductive vegetation*). Land cover classifications are very sparse in their representation of high alpine landscapes, considering for instance that in 2009 alone more than 650,000 tourists visited the *Jungfraujoch*, which is roughly located in the *Aletschhorn* and *Finsteraarhorn* region.

Furthermore, it appears that many of the terms used in CORINE and Arealstatistik only exist in the classification schemas, but not in (written) natural language. An example is *Normalwald*, as used in the Arealstatistik. The Google search engine only retrieves 23600 hits for *Normalwald*, mostly documents related to the Arealstatistik, whereas the more common equivalent *Wald*, as we find it in the spatial folksonomy, gains some 85,000,000 hits[41]. The fact that many of the terms used in land cover classifications are not represented in natural language, and thus do not link to everyday communication, can be considered a limitation of the applicability of land cover classifications to certain use cases.

The terminology of land cover classifications will only sparsely overlap with the terminology used by local people in order to refer to their surrounding environment. However, such local terminology would be crucial in order to provide information retrieval services that can cope with local affordances, as described by White and Buscher (2012). The spatial folksonomy can be considered a means for linking terms in land cover classifications with natural features that are used to describe local landscapes.

[41] Numbers correspond with the information as given by using the Google search engine, 17.06.2013

Chapter 6 Discussion

The general research question that guided this thesis is:

> *How can vagueness and ambiguity present in unstructured descriptions of natural landscapes be captured such that geographic queries can be effectively resolved (for lay communities)?*

At an early stage we decided to answer this question by dividing it into two major objectives, namely a first objective where we aimed to link landscape descriptions to space and a second objective on retrieving landscape concepts from these georeferenced descriptions. The two objectives are sketched in Figure 61 (modified from Figure 13).

Figure 61. Structure of the thesis as previously sketched in Figure 13. The two tasks are highlighted with grey background color.

In the introduction we simply called these two tasks the *value of geography* and the *value for geography*, and thus used the context of digital humanities and the work with large compilations of digitized text (e.g. Berry 2012) as an example. The value *of* geography reflects that geographic information can be considered as a means for imposing a first layer of interpretation on large data, whereas the value *for* geography should emphasize that large compilations of landscape descriptions, contain important information that serves for answering fundamental geographic research questions, such as the general question that guided this thesis.

The information gathered in each of the two objectives can be seen in the light of Sara Shatford's work on image indexing (Shatford 1986) (§2.1.3). Shatford states that rich descriptions of images, which she

argues are a precondition for successful indexing, consist of *specific* and *generic* information. Thus, we firstly retrieve specific information from landscape descriptions, in terms of toponyms grounded from text. Secondly, we capture generic landscape information, which is represented by landscape terms and their spatial use in Switzerland.

The two objectives reflect the first and the second research questions, as posed in the introduction (§1.1). The third research question is on the improvements that are introduced to information retrieval through answering research questions one and two. In the following discussion we will focus on the three research questions individually. Each research question will be associated with the major achievements, the insights that we gained and the some important limitations.

6.1 RQ 1: Linking natural Landscape Descriptions to Space

The first research question is on the requirements and specifications of an approach for linking natural landscape descriptions to spatial footprints. The major scientific challenge is toponym ambiguity. Research question one is mainly associated with results and investigations associated with the first objective, outlined in Figure 61. In the following we list and discuss all major achievements and insights gained.

6.1.1 Achievements

GGD. We introduce a new approach for linking landscape descriptions to spatial footprints, called GGD (geometric geomorphometric disambiguation, §4.2.2). The introduction of GGD is motivated by a research gap that we resolved from literature on geographic information retrieval (GIR) and on performing geoparsing and toponym disambiguation in particular. GGD is based on two heuristics, Euclidean distance and topographic similarity. We thus assume that toponyms that co-occur proximate in text are supposed to be either proximate in Euclidean space (as it is a state of the art assumption), or similar according to topographic characteristics (e.g. slope, curvature or texture). Both assumptions reflect Tobler's first law of geography, saying that everything is related, but near things are more related than distant things (Tobler 1970). *Near* in this context means proximate in text. The degree of *relation* is approximated by Euclidean proximity or topographic similarity. Both proximity and similarity are evaluated for the geographic and topographic scope of each description and thus relative measures. We apply our approach, of linking landscape descriptions to space, to different corpora and thus gather the

following three results (which in Figure 61 are simply represented by contour lines): A macro-map, a spatial index and an adaptive grid index.

Macro-Map. We compute a macro-map from a historic corpus of Swiss alpine landscape descriptions (Figure 37), which according to Cooper and Gregory (2011) allows additional readings of text, compared to traditional close reading, by imposing a first layer of information. Further analysis of the macro-map and the computation of spatial aggregates (Figure 39), and χ-map representations of these aggregates (Figure 40), helped us to understand the spatial focus of the corpus, how the focus might have changed over time and whether particular spatio-temporal events took place in some time periods.

Spatial Index. We can automatically compute spatial indexes for natural landscape descriptions. The spatial index is applied to individual documents and thus allows the retrieval of these documents in a spatial search engine. The spatial search engine can be used to test the accuracy of the spatial indexes and is thus a means for evaluation.

Adaptive Grid Index. We used the spatial index in combination with a spatial ranking and thus create an adaptive grid index for some 10,000 landscape descriptions. For each cell of the adaptive grid index we thus retrieve a list of relevant landscape descriptions (Figure 41). The adaptive grid index covers all of Switzerland, whereas regions that are described in great detail and by numerous descriptions are represented by grid cells of fine resolution. The adaptive grid index is an important source of information for retrieving local landscape information.

6.1.2 Insights

The unsolved challenge of linking natural landscape descriptions to spatial footprints is resolved as a research gap from literature and described in §2.3 (RG I). We quickly summarize some of the key issues associated with RG I. The GIR literature, and in particular approaches to disambiguate and thus link text to spatial footprints, is biased towards the use of text containing well-known places. Prominent examples are investigations that incorporate newspaper articles (Amitay *et al.* 2004, Martins *et al.* 2010), web pages (Purves *et al.* 2007) or Wikipedia articles (Overell and Rüger 2008). Approaches that can georeference descriptions referring to space by the use of less known place names, such as the names of natural landscape features, are clearly underrepresented (e.g. Leidner 2007).

The challenge in applying GIR to unpopulated, natural, fine grained and often unknown place names is the lack of explicit toponym knowledge. Such knowledge is used for performing toponym disambiguation (§2.2.2). State of the art disambiguation approaches and their application of toponym knowledge are

described in §2.2.3. Thus, answering the research question on how to link natural landscape descriptions to spatial footprints requires us to find new sources of toponym knowledge that could then be introduced to geoparsing.

Consequently, we introduced a new disambiguation approach, i.e. GGD, where we incorporate geomorphometric similarity as a disambiguation heuristic. We thus assume that toponyms, occurring proximate in text, should also refer to similar types of landscapes, in terms of local topography.

6.1.2.1 Contributions

The introduction of geomorphometric similarity to GIR and its wider application leads to a list of benefits and contributions:

1. Geomorphometric similarity can be computed for arbitrary locations in Switzerland. Its computation is **not constrained by the feature type of toponyms**. Geomorphometric similarity is only the second metric, introduced to geoparsing, which is independent from feature type. The first metric, which we also used in our approach, is Euclidean distance measured between toponym locations. The main reason for not only using Euclidean distance draws back to a finding of Brunner and Purves (2008). They showed that ambiguous toponyms in Switzerland (and Great Britain) are significantly autocorrelated and that the degree of toponym ambiguity is higher for toponyms not referring to populated places. Thus, geometric minimality is vulnerable to not being effective when applied to descriptions of fine spatial granularity.
2. The use of geomorphometric similarity for comparing mostly natural features in Text+Berg, is **in accordance with theoretical findings** on the nature of geographic objects. Smith and Mark (2003) for instance argue that geographic objects are attached to the earth's surface and thus, at least partly, determined by its shape.
3. We used the measure of geomorphometric similarity also uncoupled from geoparsing and could thus **contribute to research questions in linguistics and ethnophysiography**, namely on the meaning of toponyms (Derungs *et al.* 2013). Thus, we showed that by combining large gazetteers with fine grained terrain information, we have new means for contributing to theoretical debates. In particular, we contributed to a paradigm in linguistics saying that "Names identify individuals without utilizing any of their characteristics" (Coates 2006, p. 363). We could show that toponyms referring to natural features are often strongly related to particular topographic characteristics.
4. We could show that the application of GGD is not limited to only one corpus or type of description. We applied GGD to different corpora and found that the precision of the spatial

footprints is surprisingly similar (§4.2.5). Thus, **GGD is both generically applicable and robust in terms of the results**. The discussion of RQ 3 contains a more detailed demonstration of the contribution of GGD to information retrieval.

5. The application of GGD, and thus the grounding of toponyms from natural landscape descriptions, allowed us to draw maps from a large corpus (i.e. macro-maps, §4.2.3). These **macro-maps** were then used to deduce a first layer of knowledge. In the introduction we called this *the role of geography*, indicating that geography, or spatial distribution, is an important source of information for structuring large data. The approach we take in this thesis for computing a macro-map is **a contribution to the state of the art in GIR and literary GIS**. From a GIR perspective it can be regarded as a novelty that the product of geoparsing is used for purposes beyond the means of information retrieval (e.g. list of challenges in Jones and Purves 2008). Usually, the step from extracting geographic information towards the use of geographic information in order to analyze semantic contents is not undertaken. Literary GIS, on the other hand, prominently uses spatial representations of text in order to conduct content analysis. However, as shown in the literature review (§2.2.4), approaches associated with literary GIS, as for instance reported in Piatti (2008) or Cooper and Gregory (2011), perform the annotation of spatial references in text and the follow up mapping manually. Therefore, they usually only incorporate a limited number of documents.

6.1.3 Limitations and Improvements

6.1.3.1 Macro-Map

The macro-map is quite poor in terms of its semantic content. A density peak in the macro-map of Text+Berg can either be associated with the most prominent topic in the corpus, which is *mountaineering* or it is sufficiently particular, such that it can be related to an event. The relationship of density peaks with events is only rarely possible.

The *mountaineering* label might often be a correct explanation for particularities in the macro-map. However, it might also often be too imprecise in order to facilitate new insights on the content of the corpus. Therefore, a more profound examination of topics in the corpus would significantly improve the semantic content and the interpretability of macro-maps.

Potential Improvement:

We did a preliminary investigation where we used a *topic model* (i.e. latent dirichlet allocation: *lda*), as

for instance described in Adams and McKenzie (2013), in order to annotate the corpus for existing topics, before drawing the macro-map. Topic models afford specification of a fixed number of output topics, comparable for instance to *K-means* clustering. We thus clustered each description in Text+Berg into a predefined number of topics, using the *lda* algorithm, considering the statistical distribution of terms over the corpus. Details of the functionality of topic models are described in Steyvers and Griffiths (2007).

The insight gained from this investigation is that we could not retrieve sufficiently explicit information for each of the retained topics. As a consequence, the topics could not be associated with unambiguous labels. We believe that we are not the only ones who have struggled with labeling topics. Adams and McKenzie (2013) initially structured a natural language corpus into some 200 topics. In further investigations, however, they focus on a handmade selection of only 20 topics. We believe that manually selecting topics can have critical subjective impact on the results, which we thus want to avoid.

Nevertheless, we resolved one interesting finding from this preliminary investigation using a topic model. The topics have significant impact on the way landscapes are described. We compared landscape information retrieved from a set of regions, with landscape information retrieved from different topics. We thus found that topics are less related to each other than regions. We believe that this is an interesting starting point for further investigations.

6.1.3.2 Geomorphometric Similarity

Earlier in this thesis we argued that we were successful in computing geomorphometric similarity between toponym locations referring to different types of landscape features, such as mountains, hill or valleys (§4.2.1). Thus, each feature is represented by circular regions of different sizes. However, by using circular regions as approximations of feature footprints we assume that all types of landscape features are round. This assumption might be sufficiently precise for some feature types, such as mountains, fields or cities. For other feature types, however, this assumption is wrong. For instance rivers, valleys and streets have linear shapes and are thus not suitably captured using our approach.

One might argue that the computation of geomorphometric similarities requires clear-cut spatial footprints for each individual feature type. This might be true, but it is unrealistic. A first implication is that natural features are known to have vague boundaries (e.g. Smith 2007). In some investigations the vagueness of landscape features is approximated using fuzzy set theory (e.g. Fisher *et al.* 2004, Sinha and Mark 2010). In one in particular of these examples, fuzzy footprints of valleys were automatically extracted from terrain characteristics. This investigation took a whole PhD thesis for extracting the footprint of *only* one feature type (Straumann 2010). We thus consider the delineation of fuzzy footprints as too time

consuming and rely on an approximation that compares geomorphometric measurements gathered on multiple scales.

Potential Improvement:

We think that a more pragmatic extension of our approach would improve the reliability of similarity computations for non-circular features. One potential means of improvement could be to use gazetteers that contain more realistic geometric representations of toponym footprints. Younis et al. (2012) discuss an approach where they use a gazetteer in order to query DBpedia[42] for gathering toponym information, in particular representative spatial representations. Along the same line of research is the quattroshapes[43] initiative, which results in a global gazetteer of polygons. Quattroshapes is a conflation of data from foursquare[44] with additional, openly available, data sources in order to create "an authoritative source of polygons around a curated list of places. This gazetteer of non-overlapping polygons provides more relevant results than simple point geometries" (from quattroshapes.com).

We have our doubts that the *curated list of places*, or the information available from DBpedia, matches the fine spatial granularity requirements which are required in our work. However, it surely would be an interesting investigation and worth the effort to see how far we can get with fine granularity Swiss toponyms in combination with user generated contents, in order to enrich gazetteers and gather more suitable spatial representations of toponyms.

6.1.3.3 Geoparsing

One crucial assumption in our disambiguation approach is that toponyms proximate in text, are either proximate in Euclidean space or geomorphometrically similar. There might be autocorrelation between text distance and Euclidean proximity and/or geomorphometric similarity. However, this autocorrelation is not linked to a linguistic axiom, such as for instance syntax rules or grammar. At best, the correlation between text distance and proximity and/or similarity is often observable in descriptions, since it simplifies the comprehensibility of natural language.

Potential Improvement:

A detailed investigation of the properties of toponyms and their co-occurrence in text would clearly improve our understanding of the role and nature of spatial references. The investigation must consider different types of corpora. Thus, we would have the means to test which properties are shared between toponyms that occur proximately in text and if the type of corpus has any influence on the proximity-

[42] dbpedia.org
[43] quattroshapes.com
[44] de.foursquare.com

similarity relation. The investigation would also match a research gap that we mentioned earlier in this thesis, namely that most investigations on particularities of toponym ambiguity so far focused on gazetteers, rather than on the use of toponyms in written natural language, an example being Brunner and Purves (2008).

The output of GGD is dependent on a sizable list of input parameters, some of which are listed below:

- Size of the search window to indentify neighbors in text.
- Threshold values for Euclidean proximity and geomorphometric similarity.
- Combination of Euclidean proximity and geomorphometric similarity.
- Size of the gazetteer to perform toponym lookup.

Each of these parameters has an impact on the disambiguation result. We did some pretesting using different parameter settings, which led to the final combination as discussed in Algorithm 2. However, we did not perform a detailed investigation on the impact of each individual parameter.

Potential Improvement:
In order to carry out a detailed evaluation on the influence of individual parameter settings, we suggest using different configurations of GGD for retrieving spatial footprints from the HIKR corpus (§3.2.2). HIKR articles are associated with metadata, which can be used for selecting the optimal parameters setting, by performing a Monte Carlo analysis (e.g. Fisher 1991). An interesting outcome of such an investigation could be that the optimal parameter setting depends on the location, such that some regions are described in more detail and are thus more dependent on extensive gazetteer data or that the descriptions of some locations require a different interpretation for the proximity-similarity assumption.

6.2 RQ 2: Capturing Local Landscape Concepts from Descriptions

The second research question is on the methodological requirements for capturing landscape information from natural landscape descriptions. The major challenge is vagueness as omnipresent in natural language, and vagueness of natural features in particular. RQ 2 reflects the second objective that we set out in Figure 61.

6.2.1 Achievements

Spatial Folksonomy. The motivation for conducting an investigation on local landscape information from digitized books is based on a research gap which we resolved from ethnophysiographic literature. In our approach we incorporate a large corpus of landscape descriptions, namely Text+Berg, consisting of 150 yearbooks, reaching back to 1864. We investigated the georeferenced descriptions from Text+Berg for the occurrence of natural feature terms, such as mountain or valley. The local distribution of natural feature terms we called a spatial folksonomy (Figure 61, simplified as a tag cloud for a particular region). The spatial folksonomy covers all of Switzerland with detailed local landscape information, with a clear focus on the Swiss Alps.

Natural Features. The natural features that are *measured* for populating the spatial folksonomy are retrieved through manual annotation. The resulting list of features was compared to findings from empirical investigations on landscape concepts, with the result that some natural features overlap with basic level geographic concepts resolved from these investigations, such as *mountain* or *valley* (Table 4). However, many natural features refer to fine spatial granularity features and thus indicate that the descriptions in Text+Berg contain detailed information on mountain landscapes. Examples of fine grained features are ridge, crevasse or notch (Appendix B).

6.2.2 Insights

In the following paragraphs we aim to quickly refresh the motivation for tackling RQ 2 and the associated research gap. The description and perception of local landscapes is important geographic information, required for numerous applications, for instance in resource or land use management. In the introduction of this thesis we linked the geographic need for human sourced local information with recent work in digital humanities and *culturomics* (e.g. Michel *et al.* 2011) and argued that large digital compilations of books and methodologies for extracting information from this unstructured data is of great relevance for geography. We thus called it *the role for geography*.

A recent strand of geographic research aims at retrieving local landscape descriptions from so called *user generated content*, often represented by georeferenced and tagged social media posts (e.g. Hollenstein and Purves 2010, Purves *et al.* 2011). However, such research is often biased towards structured information, such as lists of tags or georeferenced contents. A quite different body of geographic literature, often associated with ethnophysiography and psycholinguistics, aims at retrieving local landscape descriptions through ethnographic approaches, such as interviews and/or field walks (e.g. Burenhult and Levinson

2008, Mark *et al.* 2011). Such research has the potential for gathering information at considerable spatial resolutions, however, it often falls somewhat short in terms of coverage, since collecting the data is very time consuming. We considered this a research gap (§2.3, RG III).

A second research gap that is affected by RQ 2 is the type of information that is contained in natural landscape descriptions. Ethnophysiography emphasizes the significant local variation of landscape concepts (e.g. Turk *et al.* 2011). Local variation of landscape concepts is often considered an uncertainty and associated with the two concepts *ambiguity* (§2.2.2) and *vagueness* (§2.1.6). Both uncertainties are successfully resolved in communication. However, they must be considered fundamental challenges for the interoperability of information systems, such as a GIS (e.g. Egenhofer and Mark 1995) and for the introduction of formal data structures (§2.1.7 and RG IV).

For contributing to RG III and for showing how geographic research questions could benefit from recent trends in digital humanities, we used a large compilation of digitized landscape descriptions in order to extract and structure the contained landscape information. We clearly agree that a written description is not the same as information gained in interviews or field walks, and that an author writing about a landscape might not be considered a local in the same sense as an inhabitant of a certain place. However, we still argue that landscape descriptions contain observations of sometimes considerably fine spatial granularity. We further argue that this information reflects local human landscape concepts, and is thus well suited for extending the spatial coverage of state of the art ethnophysiographic inquiries. As of a corpus we decided to use *Text+Berg* (§3.2.1), which consists of a set of 150 digitized yearbooks from the *Swiss Alpine Club*.

As a contribution to RG IV we introduced the spatial folksonomy. The spatial folksonomy is a data structure for local landscape information. For this reason we combined the adaptive grid index, resulting from the first investigation of this thesis (Chapter 4), with a set of annotated natural features (§5.2.1,

Appendix B) and computed the spatial folksonomy, which is the relative distribution of natural features for each cell of the adaptive grid. The frequency distribution of natural features is considered as local landscape information. We call it a folksonomy since the retrieved information reflects *folk* concepts (e.g. Gruber 2007b), contrasted by the use of ontology, where mostly sound and complete expert knowledge is structured (e.g. Guarino 1998).

6.2.2.1 Contributions

The particular contributions made by using large digital corpora and the spatial folksonomy for retrieving and structuring landscape information are the following:

1. The spatial folksonomy contains detailed landscape information for most of the Swiss alpine belt, which covers approximately 15,000km^2, four languages and some hundred valleys, many with local population. The **extent of the spatial folksonomy is considerably larger and more diverse, compared to the locations under consideration in ethnophysiographic investigations**. The Navajo study, for example, which is considered one of the more extensive ethnophysiographic inquiries, consists of field interviews at 18 localities (Topaha 2011). For this reason, the spatial folksonomy can be considered a clear contribution to the state of the art in ethnographic landscape investigations (RG III). The large spatial coverage, however, comes at the cost of lower level of detail and a limited signal-noise ratio. **The information gathered in our investigation is less detailed and reliable, compared to information gathered in interviews or field walks**. Therefore, the combination of ethnophysiographic findings, with landscape concepts gathered from digitized text descriptions, could lead to both large spatial extents and high level of detail.
2. We presented an approach for storing landscape information and thus referred to the *folksonomy* literature, which poses fewer methodological constrains and is more flexible for capturing vagueness and ambiguity. Both types of uncertainties are not resolved but still contained in the spatial folksonomy. Thus, the occurrence of natural features at different locations can have different meaning. This might be a problem for tasks where simply the occurrence of a particular natural feature at a particular location is measured. However, we discussed several examples of how the whole set of natural features can be used to deduce meaning (e.g. Figure 53). We argue that **by considering all natural features, uncertainties such as vagueness and ambiguity are successfully resolved by the co-occurrence of features.** Thus, the spatial folksonomy is a contribution to the debate on how to structure vague landscape information, as described in RG IV.

3. The spatial folksonomy is generated from written natural language, where both semantic and spatial information is unstructured and has first to be resolved. This is a contribution to state of the art approaches that aim to retrieve landscape information from digital data. A number of approaches use user generated content, where geographic information is often explicitly contained and the semantics are gathered from tags (Serdyukov *et al.* 2009, Wing and Baldridge 2011). However, tags are far more structured, compared to written language. Purves et al. (2011) published work where place-related information is gathered from georeferenced images, each described by some sentences. In a nutshell, there is some work that uses natural language in order to deduce local information from landscape descriptions. However, **the resolution of semantic and spatial information from unstructured text is new** and thus a contribution to the state of the art.

6.2.3 Limitations and Improvements

6.2.3.1 Natural Features and Landscape Characteristics

The list of 94 natural features, as annotated from Text+Berg, contains ambiguous cases. Some natural features have at least one alternative meaning that does not refer to a natural landscape context. A striking example is *wand* (*house wall* or *rock wall*), which by most annotators was considered as a natural feature. Ambiguities between natural feature terms and generic meanings of the same terms are sometimes introduced by the use of a simple list of nouns in the annotation task. This out-of-context use of nouns, in contrast to nouns embedded in the original sentences, often lacks the information which is needed for disambiguation. We tried to minimize ambiguous cases by only considering nouns which were annotated by at least three out of four annotators. However, once a noun is annotated as a natural feature (e.g. *wand*), we have no means to disambiguate individual occurrences of this word in follow up investigations, as either referring to a generic noun or a natural feature.

Potential Improvement:

In order to have better control over ambiguous cases, we suggest asking a larger number of annotators and to apply a hierarchical annotation process. In a first annotation task we might ask a group of at least 10 annotators to perform annotation as described in this thesis (§5.2.1). Through the incorporation of a larger number of annotators we have richer information for evaluating which nouns are particularly prone to ambiguity. The ambiguous nouns could then be tested in a second annotation task, where, instead of individual nouns, whole sentences containing the ambiguous nouns are presented to the annotators. We

thus gain useful information, for on-the-fly disambiguation during the follow up information retrieval task and finally retain more reliable landscape descriptions.

We compared the vocabulary of natural features from Text+Berg with results from empirical investigations. We thus concluded that there is overlap between the list of natural features and well-known basic levels. This conclusion might be true. However, it is important to keep in mind that we compared apples with oranges. The frequency of nouns in text is associated with less reliable semantic information, compared to the result from empirical experiments, where participants were explicitly asked to list *geographic objects* or *features*.

Potential Improvement:

In order to gain a deeper insight on the meaning of frequency of natural feature in text, and for testing if term frequency can be compared to results from empirical investigations, we suggest conducting a detailed inquiry, where we manually annotate a number of documents. Annotators are for instance asked to read these documents and list landscape features, which are particularly representative for the content of the respective text. Comparing natural feature frequencies with the set or representative features, could guide our understanding of the relation between the occurrence of landscape features in text and the semantic content of a description. Thus, we might resolve results that are crucial for all follow up investigations that deduce meaning from term frequencies (i.e. bag of word approaches).

6.2.3.2 Spatial Folksonomy

Natural features are building blocks of the spatial folksonomy. We argued that local frequency of these features can be used to deduce landscape descriptions. This might be true and clearly represents a *bag of words* approach as often applied in information retrieval (e.g. Chowdhury 2010). However, we know from our experience and linguistic theory that nouns only partly reflect the meaning of a description. The two nouns *mountain* and *house* can for instance occur in sentences as different as *The mountain destroyed my house*, and, *From my house I see a beautiful mountain*. In order to capture the meaning (e.g. role, affordance or connotation) of the natural feature *mountain* from both sentences correctly, we definitely need to gather more detailed information than just occurrence and associated counts.

Potential Improvement:

Tversky and Hemenway (1983) aimed at resolving a "taxonomy of environmental categories from perception of attributes and activities of behavior settings, and from communication about them" (p.4). They show pictures of different environmental scenes and ask participants to describe the respective *scene*

category by associating *attributes*, *activities* and *parts*. This information was then used to structure a taxonomy of scene categories.

An adoption of Tversky and Hemenway's (1983) approach is published by Kuhn (2001), where he used verbs occurring in text, in order to associate traffic objects with affordances and thus create an ontology. Purves et al. (2011) conducted an annotation task with nouns from Flickr[45] and Geograph[46], which they classified as *elements* (i.e. parts), *activities* or *qualities*. The co-occurrence of these three categories was then used to describe large spatial extents.

Applied to our approach, we could consider the natural features to be what Tversky and Hemenway (1983) called scene categories. The attributes, activities and parts associated with natural features could then be represented by adjectives, verbs and nouns, in combination with natural features in text. This could be a potential means for either building a taxonomy (Tversky and Hemenway 1983), an ontology (Kuhn 2001) or simply deducing meaningful geographic descriptions (Purves *et al.* 2011). In any case, more detailed information, additional to the distribution of natural features, would be preferable.

We did some preliminary testing and associated natural features with co-occurring adjectives, verbs and other natural features. The distribution of co-occurrences between adjectives and natural features turned out to be very broad, such that a large set of adjectives is used to describe individual natural features. This results in only limited overlap between the descriptions of the qualities of natural features and large inner feature variation. This effect could be diminished by relying on more sophisticated linguistic information, for instance by using a taxonomy of adjectives, where adjectives are classified into groups with similar connotation. By associating verbs with natural features we encountered the opposite problem. We found that only a few verbs occur frequently, such that all natural features are associated with the same activities. Here the use of a controlled subset of verbs could help to allow more detailed insights. The co-occurrence of natural features with other natural features is an exception to the two above discussed cases, in such that we were quite successful and results look promising. We gathered the co-occurrence of natural features in all cells of the adaptive grid and can thus potentially deduce different meanings of the same feature at different locations. This could be an interesting contribution to the debate of vagueness of geographic features as described in §2.1.7.

6.2.3.3 Multiple Languages

Switzerland has four languages. All of these languages, with the exception of Romantsch, are represented in the Text+Berg and HIKR corpus. In our investigations, however, we only considered German articles. This is not so much a critical issue, but clearly a missed opportunity, since in ethnophysiography

[45] www.flickr.com
[46] www.geograph.org.uk

language is often considered a major player when it comes to variation in landscape descriptions (e.g. Mark *et al.* 2007).

Potential Improvement:

An obvious improvement is to incorporate the two languages that are not covered in the spatial folksonomy yet, namely *Italian* and *French*. As a consequence we have to replicate our natural feature annotation task for these languages.

The products, which are spatial folksonomies in three languages that cover the same region, could be considered a unique data basis for investigating the impact of written language on landscape perception. The Text+Berg corpus is particular interesting. Early yearbooks contain articles in one of the three languages, depending on the mother tongue of the author, whereas in more recent yearbooks all articles are translated, such that each article is available in three languages. From this we could investigate if landscape descriptions are more authentic if they are written in a particular language, compared to articles that are translated.

The multi-language extension of our investigation would introduce several new uncertainties. One uncertainty is that the influence of language overlaps with the influence of change over time of perception and nature. It could prove to be very complex to distinguish the different influences on variation. Another uncertainty is related to translation problems. Natural feature terms used in different languages have sometimes no, or only limited, semantic overlap such that translation is problematic. This could hinder us from comparing landscape descriptions retrieved for different languages.

6.3 RQ 3: Improving Information Retrieval

The third research question focuses on the improvement introduced to information retrieval through the two objectives sketched in Figure 61.

6.3.1 Achievements

Geoparsing Evaluation. The geoparsing algorithm GGD was used to compute a spatial index for two corpora. From the spatial index we designed a spatial search engine that allows for evaluating the geoparsing algorithm (§4.3.1.1), as well as the effectiveness of GIR in general (§4.3.1.2). We could thus show that GGD outperforms other state of the art GIR approaches for a fine granularity corpus (Figure

31), and that traditional information retrieval cannot achieve the precision of GGD, if compared on a corpus of outdoor descriptions and by using queries of fine spatial resolution (Figure 33).

Landscape Comparison. We used the spatial folksonomy in order to compare the description of different landscapes in Switzerland. The comparison between different landscapes suggests that the landscape information stored in the spatial folksonomy can uncover expected relationships between different landscapes, by conducting qualitative (Figure 52) and quantitative (Figure 53) examinations.

Explaining Landscape Variation. The spatial folksonomy has broad spatial coverage and a depth of semantic content, such that it can be used for statistical hypothesis testing. We thus compared variation in descriptions to variation of topography, as deduced from geomorphologic classifications, and could show that the characteristics of descriptions are clearly related to local topography (Figure 54). Thus, local landscape descriptions could be partly deduced from the shape of the earth's surface.

Land Cover Classifications. By comparing the spatial folksonomy to official land cover classifications we identify different spatial coverage and different semantic content (Figure 59 and Figure 60). Both particularities suggest that the land cover classifications and the spatial folksonomy could be fruitfully combined for tackling further research questions on the nature of landscape descriptions and its applicability to fine granularity GIR.

6.3.2 Insights

The goal of this thesis is to retrieve local information from written landscape descriptions. Under the umbrella of the first two research questions we discussed the major insights gained from resolving fine granularity *specific* (i.e. toponyms, RQ 1) and *generic* (i.e. natural features, RQ 2) information from landscape descriptions (c.f. Shatford 1986) (§2.1.3). The product is called a spatial folksonomy. RQ 3 is on the improvements introduced to information retrieval by applying this information. We thus briefly recap the role of local information in information retrieval.

White and Buscher (2012) from Microsoft esearch stated that local knowledge is the key for knowing local interests which has "implications for search and recommendation systems" (p.1607). One implication mentioned by White and Buscher (2012) is that local interest varies, with the consequence that knowing local interest is crucial to, for instance, suggesting a restaurant. This is a compound of the *ethnophysiographic hypothesis*, as presented earlier in this thesis (Mark *et al.* 2007), and the debate on *naïve geographical knowledge*, introduced by Egenhofer and Mark (1995), both applied to an information retrieval context. With the ethnophysiographic hypothesis it is stated that people from different cultures

and language groups use differing concepts for referring to the local environment. Naïve geographical knowledge, on the other hand, emphasizes the importance of knowing lay people's understanding of the geographical world, in order to design useful applications.

We will separately discuss the specific information retrieved in Chapter 4 and the combination of specific and generic information as resulted from Chapter 5 (i.e. spatial folksonomy). Specific information, in terms of spatial footprints resolved from landscape descriptions is compared to the state of the art in *geographic information retrieval* (GIR), whereas applications of the spatial folksonomy are discussed for potential means of contributing to the debate of local information in *information retrieval* (IR).

6.3.2.1 Contributions to Geographic Information Retrieval.

As discussed in the context of RQ 1 we introduced a new approach for linking natural landscape descriptions to spatial footprints, i.e. GGD. GGD incorporates local topographic subtleties in order to guide the toponym disambiguation process (§4.2.1). We applied GGD to two corpora, namely *Text+Berg* (§3.2.1) and *HIKR* (§3.2.2), both describing landscapes in fine spatial granularity.

1. The spatial footprints resolved by applying GGD to Text+Berg were compared with a simple GIR baseline. Since the corpus is not associated with ground truth information, we conducted a user centered evaluation, where we asked a group of experts to judge the precision of document retrieval. We gained relevance judgments for a set of 10 spatial queries (§4.2.5.1). The results clearly indicate that **GGD outperforms a simple disambiguation baseline** (spatial precision: 0.8, Figure 31 and Figure 32). The evaluation is of quite small extent and the compared GIR baseline is relatively simple. However, this does not change the fact that we gained a considerably high spatial precision for a corpus that must be considered a challenging touchstone for GIR (e.g. Leidner 2007). The spatial precision of 0.8 could for instance be compared to spatial precision as published in Purves et al. (2007), where the mean spatial precision for some 38 queries (and two annotators) is 0.5 (Figure 8).
2. The HIKR corpus allowed for conducting an **extensive evaluation**, since each HIKR article is associated with metadata on topic and way points. The metadata is used in an automatic evaluation process where we tested some **5000 geographic queries**, each consisting of spatial and topical information. On the downside, metadata is not equal to a gold standard, such that it is not guaranteed that the topic and the way points, as represented in the metadata, are explicitly mentioned in the description. Nevertheless, **GGD clearly outperforms the two IR baseline approaches, both based on string search**. The best spatial precision of GGD is 0.87 (Figure 34, 5km), the best precision for queries containing spatial and topical information is 0.73 (Figure 33,

1km). The precision values of GGD are not only significantly higher than the precision retrieved by the string search baseline. The precision values are also surprisingly high compared to other evaluation initiatives in GIR, such as GeoCLEF, which was the most extensive GIR evaluation (e.g. Mandl *et al.* 2008). In GeoCLEF they could not find indications that GIR can outperform IR for queries containing spatial information, which might relate to the relatively simple spatial queries that were incorporated, mostly aiming for city or country names (e.g. Kornai 2006). We found that for fine spatial granularities in the queries, as well as the data, state of the art IR cannot compete with GIR. This is a contribution to the state of the art in GIR, since previous improvements on classical IR were only achieved by incorporating complex spatial relations in queries, such as distance or directions (e.g. Purves *et al.* 2007). Thus, the fine spatial granularity of GGD offered the possibility for testing GIR in *continuous* space (i.e. thousands of queries continuously distributed over broad spatial extents). **The results suggest that for resolving queries of fine spatial granularity with sufficient precision, the incorporation of geographic intelligence in IR is indispensable**.

3. The impact of **buffer size**, associated with spatial queries, was evaluated separately. Different buffer sizes can be related to different human information needs and affordances. Thus, the precision per buffer size is an important predictor for how important the incorporation of geographic information is in different contexts. GGD outperforms the IR baselines for all evaluated buffer sizes (1, 2, 5 and 10km) (Figure 33). However, the relative difference between the performances is maximal for buffer size 10km. Consequently, **the incorporation of geographic information in IR is more relevant if the information need covers broad spatial extents**. To our knowledge, this effect has never been shown so far. We thus consider it a contribution to the state of the art in GIR.
4. A last finding from applying GGD to GIR is not related to a contribution, but rather considered an indication for the *generic* applicability of GGD. We applied GGD to two corpora, both containing natural landscape descriptions. Text+Berg mostly contains documents of approximately five to six pages in length. HIKR, on the other hand, consist of short reports of outdoor activities, related to different topics. The application of GGD to both corpora resulted in comparable precision values. This indicates that **GGD is generic and robust**. Generic in terms of being applicable to different corpora describing natural landscapes. Robust since the retrieved results are comparable. This is an important finding for all future applications of GGD.

6.3.2.2 Contributions to Local Landscape Concepts in IR

We did not apply the spatial folksonomy to an information retrieval task, such that we could make a statement on its performance in terms of precision or recall, as discussed above, on the example of GGD. However, we compared the local landscape information in the spatial folksonomy to numerous other landscape descriptions of different scale and nature, and can thus draw conclusions on the characteristics and applications of the information, stored in the spatial folksonomy.

1. We compared the focus of the spatial folksonomy with official land cover classifications, namely *Arealstatistik* (§3.4.1) and *CORINE* (§3.4.2). The two land cover classifications use classification schemas that mainly apply to populated or agricultural areas (Figure 58). This is for instance reflected by the large number of classes available for describing settlements. The Swiss mountains, on the other hand, are largely classified as being either of the type *glacier* or *rock* (Figure 60). This contrasts with the focus of the spatial folksonomy, which contains rich descriptions of the Swiss Alps and only sparse information for most locations in the *Swiss Mittelland*. Thus, the **spatial folksonomy and the land cover classifications are complementary in terms of spatial focus**, such that a combination of the two sources of landscape information would extend the spatial coverage.
2. The schema of classes, used in land cover classifications is accompanied with detailed definitions and application rules for each available class. This guarantees that each class is homogenously used in the whole of Switzerland or the whole of Europe, which is an important characteristic in the debate on interoperability of geographic information (e.g. Bishr 1998). However, application rules can be the source of artifacts. Many class names do not reflect terms that are of wider use in natural language, but have been specially introduced for the purpose of classification. In addition, the homogeneous application of classes over large spatial extents contradicts with the nature of geographic features. It is widely accepted that the meaning of geographic features differs over space (e.g. Burenhult and Levinson 2008). These artifacts, mainly caused by the artificial character of the taxonomies used in land cover classifications, are resolved in the spatial folksonomy. The spatial folksonomy only contains **natural features that are well represented in written natural language and thus actively used in communication**. **Additionally, the use of natural features in text is not constrained in meaning**. The vagueness associated with geographic features is retained in the combination of natural features that is used to describe a particular region. Similar to the finding from comparing the spatial coverage, we suggest accepting the two landscape descriptions as being complementary. A **combination of the**

information in the spatial folksonomy and land cover classifications could lead to a balance between interoperability and local description that reflects active language use.

3. We showed that by clustering the spatial folksonomy we create groups of similarly described regions, or types of landscapes (§5.2.3.2). Types of landscapes, or landscape typologies are a well-known data source in spatial planning (e.g. Mücher *et al.* 2010). Usually landscape typologies, such as the typology of Swiss landscapes as used as a reference in this thesis (§3.4.3), are computed through an aggregation of information from land cover classifications and other physical layers, such as population density or landscape inventories (e.g. Van Eetvelde and Antrop 2009). The computed typology can have crucial impact on political decision making processes, such that some types of landscapes are preserved, whereas others are more extensively used. We did not find many similarities between our landscape typology, computed from the spatial folksonomy, and the official typology for Switzerland (§3.4.3). We could therefore not use the typology in order to evaluate our approach. However, we found clear **indication that if people describe landscapes, they come up with a typology that is very different from typologies that are deduced from expert knowledge**. We thus argue that our approach shows a potential way for **automatically generating alternative landscape typologies that are folk-oriented**, and thus shed light on how people perceive the world. This might also offer new means for allowing people **more fundamental participation in decision making processes on landscape relevant scales**, as it is for instance claimed by the *European Landscape Conservation*:

   > *"Landscape exists because it is visible. A landscape policy which involved only experts and administrators, who themselves are often specialists, would result in landscapes that were imposed on the public, just as in the days when landscape was produced by and for an elite."*[47]

4. **Natural features**, as represented in the spatial folksonomy, were explicitly selected through a rule-based annotation task. This guarantees that **natural features explicitly refer to landscapes**, reflecting the notion that landscapes are wholes consisting of parts (Naveh and Lieberman 1984), in our case represented by natural features. A comparable approach of using controlled lists of place-related terms was undertaken by Purves et al. (2011). They aimed to describe places from user generated content (i.e. *place-related facets*) and argue that controlled lists of terms provide a rich "basis for analysis and discrimination". We clearly agree. There are a number of recent GIR studies where mostly unfiltered, frequently occurring terms are used as place-related information

[47] p.28, www.coe.int/t/dg4/cultureheritage/heritage/Landscape/Publications/PaysageDeveloppement%20_en.pdf

(e.g. Serdyukov *et al.* 2009, Wing and Baldridge 2011). This information might be sufficiently rich for supporting automatic geocoding, which is the goal of both investigations. However, **the use of unfiltered sets of frequently used terms in order to characterize place, in our view, leads to semantically poor and often unspecific or confusing descriptions**.

5. The broad spatial coverage and the rich semantic information in the spatial folksonomy allows for moving from observing and testing differences between landscape information, towards investigations on potential explanations for variation. Along this line of argument we explored whether the differences between landscape information in the spatial folksonomy correlates with other geographic variables, such as topographic characteristics. Interestingly, the pattern evolving from similarities between landscape information significantly correlates with topographic characteristics (0.43, Figure 54). Thus, **topography seems to be an important driver of landscape descriptions in the Swiss Alps**. This overlaps with a finding of Gschwend and Purves (2012), where they found indication that certain tags in user generated content are predominantly used for describing particular topographies. Such statistical inquiries are only possible due to broad spatial coverage and detailed landscape information. We are aware that investigations on correlations between human concepts and physical measurements are prone to the *ecological fallacy*, the *minimum areal unit problem* and *spatial autocorrelation* (as discussed in O'Sullivan and Unwin 2003). At the same time, such investigations reflect recent trends in linguistics, where structural variation in language is explained through geographic variables (e.g. Everett 2013). We see two main applications emerging from the deduction of human concepts from physical measurements. Firstly, this could be the starting point of a **deep understanding of why people describe landscape the way they do**. Secondly, the dependency between landscape information and physical factors could serve as a means for **deducing unknown local landscape information from available physical measurements**, and thus significantly tackle the lack of local knowledge in information retrieval as elaborated by White and Buscher (2012).

6.3.3 Limitations and Improvements

6.3.3.1 Applicability of GGD

We applied GGD to two different corpora and thus argued that the approach is generically applicable. This argument holds true under two conditions. Firstly, GGD was designed to geoparse natural landscape descriptions and was only applied to corpus data of this type. Secondly, the focus on geomorphometric similarity requires the described spatial extents to be geomorphometrically diverse. In cases where

topography is not an important characteristic of the described landscape GGD will probably not be suitable for correctly resolving toponym ambiguity.

Potential Improvement:

By incorporating geomorphometric characteristics in geoparsing we could show that spatially continuous information is useful for characterizing toponym locations, such that the disambiguation precision is improved. Geomorphometric information can be considered as only one example of continuous information. The methodological approach taken in GGD can be used similarity for other types of spatially continuous information, such as temperature and income distributions or colors and textures retrieved from satellite imagery. Depending on the context of the corpus such information could be useful in order to support toponym disambiguation.

6.3.3.2 Evaluation of the Spatial Folksonomy

The landscape information stored in the spatial folksonomy was used in qualitative and quantitative comparisons, and it was related to the content of land cover classifications. All comparisons suggested that the content of the spatial folksonomy meets our expectations and that the information, particularly in high alpine regions, is considerably detailed. However, we did not compare the landscape information in the folksonomy with ground truth information and can thus not finally conclude on the reliability and, importantly, on the level of detail of the information.

Potential Improvement:

The spatial folksonomy could be accompanied by an ethnographic investigation, where local people in the Swiss Alps are asked for their landscape concepts. Information from the ethnographic inquiry could shed light on further applications of the spatial folksonomy. If local landscape concepts overlap with the landscape information in the spatial folksonomy, the spatial folksonomy could be applied to guide local decision making processes, such as way finding in natural landscapes or spatial planning on community level.

It is likely that the information retrieval process, as described in this thesis, is to be extended, such that more detailed context information from text can be retrieved. Some suggestions on how to improve the geographic information retrieval are outlined in §6.2.3.2.

6.4 Synthesis

The general question that guided this thesis was:

> *How can vagueness and ambiguity present in unstructured descriptions of natural landscapes be captured such that geographic queries can be effectively resolved (for lay communities)?*

What is the reason for posing this general research question and why is it important to have the means for answering it? We started this thesis by discussing the role *of* and *for* geography in the context of the availability of large digital text corpora and thus mentioned the so called *data avalanche* (Miller 2010) and, as a consequence, the *great unread* (Cohen 1999). The two emerging questions are: *How can we get a first overview on large data?* And *How can we retrieve detailed information from large digitized compilations of text, such that we can make alternative contributions to fundamental research questions?* We argue that mapping is a powerful tool for gathering a first impression on large data, this we consider the role *of* geography. On the other hand, we argue that the information contained in large digitized text compilations is crucial for making an alternative contribution to geographic research, for instance on how people perceive and describe local landscapes. This we called the role of information in digitized descriptions *for* geography. Both these aspects are reflected in the above general research question and thus make an argument for its significance.

But, since the general research question is of utterly importance, did we also find suitable means for answering it? We showed that the answer to the general research question is twofold. Firstly we need a new approach for linking unstructured natural landscape descriptions to spatial footprints and, secondly, a geographic information retrieval approach that retrieves specific, local landscape information from georeferenced text. These two consecutive objectives are each associated with one fundamental geographic uncertainty. These uncertainties, namely ambiguity and vagueness, are both listed in the general research question and have to be tackled separately.

Ambiguity. We discussed in detail that ambiguity of place names is a fundamental problem for georeferencing text and we thus introduced and evaluated a new approach that resolves ambiguity even on very fine spatial granularity. This was not possible before and therefore we would argue that we successfully dealt with ambiguity and that we contributed to the state of the art in the related domains. The first uncertainty listed in the general research question can thus be considered as successfully resolved.

Vagueness. Theoretical findings on vagueness of geographic information were described in the state of the art chapter and then used as a motivation for not relying on formal knowledge structures for storing

landscape information. The landscape information that we retrieve with our approach reflects landscape descriptions as given by a large number of people and is thus expected to contain many vague cases, such as differing or even contradicting landscape concepts. For this we coined the phrase *linguistic vagueness*. We preserved linguistic vagueness in our final data structure on landscapes. The list of landscape terms that is stored and used for describing individual landscapes is allowed to contain synonyms and contradictions. Spatial vagueness, on the other hand, was mostly ignored in this thesis. The spatial folksonomy maps the use of landscape terms to spatial regions (i.e. grid cells). However, we did not resolve individual footprints of landscape features and, more importantly, we did not use our retrieved information in order to significantly contribute to the debate on the nature of vagueness. We designed and populated a spatial folksonomy in order to deal with vagueness in landscape descriptions, but we can still not make a statement on the role of vagueness in descriptions of Swiss mountain landscapes, as the general research question would imply.

One last question that is relevant in the context of discussing the contributions from a broad synoptic perspective is how the two objectives, set out in the general research question, contribute to one complete picture. The answer to this question can be sharp. The title of this thesis is *From Text to Landscape*. This is a sharp summary of the overall goal. We want to go all the way from natural landscape descriptions in text to spatially and semantically explicit information on landscapes. As we have indicated above, this process is only possible by combining the two geographic roles or, more precisely, by combining the two objectives of this thesis, i.e. the linking of landscape descriptions to spatial footprints and the retrieval of local landscape information.

Chapter 7 Conclusion

This chapter concludes the thesis by listing the key findings and by discussing two future work projects that significantly extend the work presented in this thesis.

7.1 Findings

In the discussion we aimed at answering the three research questions that guided this thesis. In this chapter, which is on the final conclusion, we will list the key findings for each of four central topics:

1. Automatic macro-mapping
2. Linking natural landscape descriptions to spatial footprints
3. Characterizing landscapes using text descriptions
4. Storing landscape information in a Spatial Folksonomy

The four topics that constitute the structure of the conclusions are related to the two objectives of this thesis as sketched in Figure 13. Additionally, each of the four topics could be considered a label of one of the research gaps, as described in §2.3.

7.1.1 Automatic Macro-Mapping of a Corpus of natural landscape descriptions

Automatic macro-mapping was resolved as a research gap from the literary GIS literature and relates to the state of the art in literary GIS of manually annotating text documents in order to generate maps from books (RGI). The following findings are part of the publication Derungs and Purves (2013).

- **Swiss Alpine Map.** We introduced an approach for automatic macro-mapping of natural landscape descriptions and thus computed a macro-map of the Text+Berg corpus that shows an intuitive footprint of the Swiss Alps, or Swiss alpine activities, with density peaks in the *Bernese*

Oberland, the Valais, the *Haute Savoy* and the *Bernina* region. The Text+Berg macro-map is computed from 10,000 text documents, distributed over the whole of Switzerland and the last 150 years.

- **Change Over Time.** The segregation of the macro-map into 20 year intervals helps to identify regions that are continuously covered with descriptions throughout the entire 150 years timespan. These regions are ideally suited for investigating the change of landscape descriptions over time.
- **Event Detection.** The representation of 20 year intervals of the macro-map as χ-maps, i.e. highlighting spatial under- and over-representation, is a powerful tool for detecting events, such as the opening of the railway connection crossing the *Albula Pass.*

7.1.2 Linking Natural Landscape Descriptions to Spatial Footprints

State of the art GIR approaches for linking text to space are design to incorporate spatial information of coarse spatial granularity. Thus, the linking of fine spatial granularity information, as for instance contained in natural landscape descriptions, is a research gap (RGII) and requires the introduction of new approaches and heuristics (Derungs *et al.* 2011, Derungs and Purves 2012, 2013).

- **Geomorphometric Similarity.** In a number of investigations we could show that geomorphometric characteristics of toponyms, in terms of slope, relief and texture, can be used to compute geomorphometric similarity. Thus, we can successfully distinguish the characteristics of different types of toponyms, such as cities, rivers and mountains (Derungs and Purves 2012). Additionally, we found indications that toponyms containing the same generic part in their name, such as *Horn* in *Matterhorn* and *Finsteraarhorn*, are often geomorphometrically similar (Derungs *et al.* 2013). The latter example shows ways of using geomorphometric measures of similarity in order to contribute to research questions from social sciences and the humanities.
- **Geometric Geomorphometric Disambiguation (GGD).** The measurement for geomorphometric similarity was combined with state of the art Euclidean distance and introduced to a new approach of geoparsing (i.e. *GGD*), particularly suited to resolve spatial footprints from natural landscape descriptions.
- **Evaluation of GGD.** GGD was applied to two different corpora, namely 10,000 articles from the Text+Berg corpus on Swiss mountain history and 26,000 articles from the HIKR homepage, describing outdoor activities in Switzerland. The product, namely spatially indexed documents, was used for designing an evaluation task, in terms of a spatial search engine (e.g. Derungs *et al.* 2012). An extensive evaluation, based on the georeferenced articles from HIKR, indicated that by

using GGD we can retrieve information of high spatial precision for spatial queries. In addition, we could show that for queries containing fine spatial granularity information, the application of GGD is significantly more effective than a state of the art string search, as used by traditional search engines.

7.1.3 Characterizing Landscapes using Text Descriptions

Landscape information is usually retrieved through either large data compilation campaigns that aim at retrieving homogeneous and interoperable information from broad spatial extents (e.g. land cover classifications, such as CORINE), or very detailed ethnographic inquiries, where individual people are asked for their landscape concepts. The retrieval of detailed, personal landscape information for broad extents from written documents is considered a research gap (RGIII). Most of the following findings are published in Derungs and Purves (2013):

- **Georeferenced Landscape Information.** The frequency of occurrences of natural feature terms in descriptions was linked with a spatial index for each description. We thus retrieved localized lists of natural feature terms. We call this local landscape information a spatial folksonomy of Swiss mountain landscapes. The spatial folksonomy allows for quantitative and qualitative comparisons between landscapes. Qualitative comparisons have shown the landscape information as stored in the spatial folksonomy is detailed, precise and easy to understand.
- **Quantitative Landscape Comparison:** Quantitative comparisons between landscapes can answer two sorts of questions. By comparing a particular landscape to all other cells in Switzerland, we can answer questions such as: *How similar is the description of the region X to all other regions in Switzerland?* By applying clustering to the spatial folksonomy, we can tackle questions such as *What different types of landscapes can be identified in Switzerland, in terms of their description?* Thus, the spatial folksonomy helps for answering new questions or for finding new answers to old questions.

7.1.4 Storing Landscape Information in a Spatial Folksonomy

Landscape information is vague and challenging to capture. We thus suggested the use of a folksonomy that captures local information from descriptions, which is an alternative to the prominent use of (formal) ontology. We considered the successful capturing of vagueness in a spatial folksonomy an individual research gap since (RGIV).

- **Folksonomy vs. Ontology.** Many approaches for structuring geographic information suggest the use of formal ontology. We decided to use a folksonomy instead, mainly since the use of folksonomy is not dependent on sound and complete information. Soundness and completeness are usually not provided for information on landscapes, where for instance local variation in landscape perception is very pronounced. Further, the use of folksonomy reflects the bottom up character of the information that we store in the spatial folksonomy, where people describe the surrounding landscape using their (written) natural language.
- **Spatial Coverage.** Compared to many ethnographic approaches, that are often also interested in local information, the spatial folksonomy has broad spatial coverage. However, this comes at the cost of reliability and level of detail. Interestingly, we found that the focus of the spatial folksonomy and land cover classifications are complementary, such that a combination of the two would clearly extend the area which is covered with detailed landscape information in Switzerland.
- **Language Use.** The natural features that are used to populate the spatial folksonomy explicitly link to natural landscapes. By contrast, the vocabulary used in official land cover classifications is most often artificially introduced and only loosely related to language use. The information collected in the spatial folksonomy could thus be considered a first step towards interoperability of local information with information as used in everyday communication. This is for instance a relevant building block for local information retrieval.
- **Explaining Variation.** The rich and detailed landscape information in the spatial folksonomy can be related to explanatory variables in order to explain its variance. We did a case study where we linked the variation in landscape information with topographic characteristics. Results indicate that local topography is a driver for local landscape descriptions. This finding is interesting for two reasons. Firstly, local knowledge plays a central role in information retrieval, but is often only available at relatively high costs. Thus, the deduction of local knowledge from existing physical parameters could clearly improve the state of the art in local search. Secondly, the spatial folksonomy constitutes a rare opportunity, namely the availability of rich localized information on natural landscapes. This is a precondition for linking landscape information to explanatory variables and thus gaining a deeper understanding on dependencies between physical parameters and human perception.

7.2 Outlook

As an outlook we describe two means for extending the work of this thesis, related to both geographic roles described in the introduction. Firstly, we describe means to extend the spatial coverage of the retrieved information. This requires new means for linking text to spatial footprints, depending on the data, and can thus be considered a contribution to the role *of* geography. Secondly, we describe an outlook on incorporating different types of text descriptions, such that the retrieved information is applicable to answer research questions from a wider range of scientific disciplines. Such information plays an important role *for* geography.

7.2.1 Extending the Spatial Coverage

The spatial coverage of landscape information as retrieved in this thesis is bound to the extent of Switzerland. It is a reasonable extension to reproduce the methodological approach of this thesis, while incorporating landscape descriptions that cover broader spatial extents.

One example would be to incorporate the historic yearbooks from the (British) *Alpine Club*[48]. From its beginning in 1857, the Alpine Club realized expeditions all over the world, such as in *the Alps*, *the Himalayas*, *the Karakorum* and in *Patagonia*. The descriptions from the Alpine Club are comparable to the data retrieved from Text+Berg. Thus, this extension would not have major influence on the methodology. However, the incorporation of descriptions with global, instead of country-wide, distribution requires the use of more extensive gazetteers, with still fine spatial resolution. On the one hand, it is challenging to compile such large and detailed gazetteers. Only a few openly available gazetteers have global coverage, of which one, Geonames[49], is presumably not fine grained enough. A second challenge constitutes the relationship between gazetteer size and degree of toponym ambiguity, as discussed in the state of the art of this thesis (§2.2.2). Thus, the incorporation of natural landscape descriptions with global focus requires an extension of the disambiguation method.

An interesting global data set that requires fundamental methodological adoptions was mentioned in the introduction. Michel et al. (2011) published an article in *Science*, starting off with the words:

> "We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. We survey the vast

[48] www.alpine-club.org.uk
[49] www.geonames.org

terrain of ‘culturomics,’ focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000.” (p.176)

In this paper it is shown that digitized and structured text can be used for resolving temporal plots that contain relevant information for analyzing cultural trends (i.e. *culturomics*). The data, stemming from some 4 million digitized books, as used by Michel et al. (2011), is provided by *Google* and freely available from the web[50]. It is distributed as *Ngrams*, which is the frequency of tokens consisting of *n* words (e.g. *Lake* is a 1gram and *Salt Lake City* a 3gram).

The Ngrams from the 4 million digitized books presumably contain interesting and relevant information on landscape descriptions, covering the last two centuries and broad extents of the globe. However, the challenge that has not been addressed by Michel et al. (2011), is to spatially index Ngrams. Ngrams are fundamentally different from natural language text, and thus require adopting the methodology of geoparsing, as introduced in this thesis. One potential means for resolving spatial footprints from Ngrams might consists of using co-occurrences of natural features and toponyms in Ngrams (e.g. *Berg Zürich*). However, toponym ambiguity and the global coverage of the data must all be considered fundamental challenges, such that it might not be possible to filter Ngrams relevant landscape information.

7.2.2 Extending the Topical Coverage

Besides the incorporation of landscape descriptions that cover broad spatial extents, and thus link to the role *of* geography, we could also extend the type of information that is incorporated and thus show that digitized descriptions can contribute information *for* answering a variety of basic geographic research questions. In this thesis we focused on corpora that described natural alpine landscapes and associated outdoor activities.

We suggest further work that extends the topical coverage of this thesis by incorporating corpus data with descriptions on different topics, and thus extending the applicability of the resolved information. The range of potentially available topics is broad. One example is descriptions of historic, local weather phenomena, such as for instance recorded by monasteries or local farmers. Information retrieved from these descriptions could be complementary to the data scarcity for past climates. Another example is historic accounts of legal decisions that bear the potential of giving detailed insights on how community life and legal issues have changed over time. The significance of such information, from a computer linguistic perspective, is recognized by Piotrowski (2012) in the book *Natural Language Processing for*

[50] storage.googleapis.com/books/ngrams/books/datasetsv2.html

Historical Texts. The corresponding corpus of digitized law texts, dating back to the year 800, is described by Höfler and Piotrowski (2011).

We could also start thinking about incorporating textual information that refers to non-geographic space. Two examples are information on the human brain and the universe. Both domains are extensively researched, information is collected in large bodies of literature and, importantly, both domains know spatial *regions* and these regions are used as spatial references in the descriptions. We would thus compute spatial folksonomies (*folk* in this case would refer to scientists, which is probably not in agreement with its original meaning) of domains that are usually not associated with spatially and semantically structured information. This could bear the potential of making, in particular old information better accessible.

An extension of the topical coverage of this thesis that would require major methodological modifications is the incorporation of topically inhomogeneous compilations of descriptions, i.e. corpora that contain several co-existing and fundamentally different topics. Thus, an important first objective would consist of automatically separating topics. The use of *topic models*, as used in a comparable context by Adams and McKenzie (2013) and briefly described in §6.1.3.1, could be one solution. However, topic models might not introduce sufficient information on the meaning of automatically resolved topics, such that other approaches of document classification must be adopted.

References

Adams, B. and McKenzie, G., 2013. Inferring thematic places from spatially referenced natural language descriptions. *In*: *Crowdsourcing Geographic Knowledge*. Springer, 201–221.

Agarwal, P., 2005. Ontological considerations in GIScience. *International Journal of Geographical Information Science*, 19 (5), 501–536.

Agirre, E. and Rigau, G., 1996. Word sense disambiguation using conceptual density. *In*: *Proceedings of the 16th conference on Computational linguistics-Volume 1*. 16–22.

Alazzawi, A.N., Abdelmoty, A.I., and Jones, C.B., 2012. What can I do there ? Towards the automatic discovery of place-related services and activities. *International Journal of Geographical Information Science*, 37–41.

Amitay, E., Har'El, N., Sivan, R., and Soffer, A., 2004. Web-a-Where : Geotagging Web Content. *In*: M. Sanderson, K. Järvelin, J. Allan, and P. Bruza, eds. *Text*. ACM, 273–280.

Andrienko, G., Andrienko, N., Mladenov, M., Mock, M., and Poelitz, C., 2010. Extracting Events from Spatial Time Series. *In*: *Proceedings of the 14th International Conference Information Visualisation*. 48–53.

Bateman, J., Hois, J., Ross, R., and Tenbrink, T., 2010. A linguistic ontology of space for natural language processing. *Artificial Intelligence*, 174 (14), 1027–1071.

Battig, W.F. and Montague, W.E., 1969. Category Norms for Verbal Items in 56 Categories: A Replication and Extension of the Connecticut Category Norms. *Journal of Experimental Psychology Monographs*, 80 (3), 1–46.

Bayardo, R.J., Ma, Y., and Srikant, R., 2007. Scaling up all pairs similarity search. *In*: *Proceedings of the 16th international conference on World Wide Web*. 131–140.

Bensalem, I. and Kholladi, M.-K., 2010. Toponym disambiguation by arborescent relationships. *Journal of Computer Science*, 6 (6), 653–659.

Berry, D.M., 2012. *Understanding Digital Humanities*. Palgrave Macmillan.

Beven, K.J. and Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Journal*, 24 (1), 43–69.

Bibby, P. and Shepherd, J., 2000. GIS, land use, and representation. *Environment and Planning B*, 27 (4), 583–598.

Bishr, Y., 1998. Overcoming the semantic and other barriers to GIS interoperability. *International Journal of Geographical Information Science*, 12 (4), 299–314.

Bittner, T., 2009. Logical properties of foundational mereogeometrical relations in bio-ontologies. *Applied Ontology*, 4 (2), 109–138.

Bittner, T., 2011. Vagueness and the trade-off between the classification and delineation of geographic regions--an ontological analysis. *International Journal of Geographical Information Science*, 25 (5), 825–850.

Bittner, T., Donnelly, M., and Smith, B., 2009. A spatio-temporal ontology for geographic information integration. *International Journal of Geographical Information Science*, 23 (6), 765–798.

Bittner, T. and Winter, S., 2004. Geo-semantics and Ontology Extended abstract. *In*: *Proceedings of the Bentley Empowered Conference, Orlando, Florida*.

Black, M., 1937. Vagueness. An exercise in logical analysis. *Philosophy of science*, 4 (4), 427–455.

Blaylock, N., Swain, B., and Allen, J., 2009. TESLA: A tool for annotating geospatial language corpora. *In*: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 45–48.

Bodenhamer, D.J., Corrigan, J., and Harris, T.M., 2010. *The spatial humanities: GIS and the future of humanities scholarship*. Bloomington: Indiana University Press.

Bohnemeyer, J., Burenhult, N., Enfield, N.J., and Levinson, S.C., 2004. Landscape Terms and Place Names elicitation guide. *Field Manual Volume 9*, 9, 75–79.

Borlund, P., 2003. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information research*, 8 (3), 3–8.

Bossard, M., Feranec, J., and Otahel, J., 2000. CORINE land cover technical guide: Addendum 2000.

Boyd, D. and Crawford, K., 2011. Six provocations for big data. *In*: *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*.

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H., 2004. TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2 (4), 597–620.

Brunner, T. and Purves, R.S., 2008. Spatial Autocorrelation and Toponym Ambiguity. *In*: *GIR '08: Proceedings of the 2nd international workshop on Geographic information retrieval*. 25–26.

Burenhult, N. and Levinson, S., 2008. Language and landscape: a cross-linguistic perspective. *Language Sciences*, 30 (2-3), 135–150.

Buscaldi, D., 2011. Approaches to Disambiguating Toponyms. *In*: R. Purves and C. Jones, eds. *Letters on Geographic Information Retrieval*. ACM Sigspatial Special, 16–20.

Buscaldi, D. and Magnini, B., 2010. Grounding toponyms in an Italian local news corpus. *In*: *Proceedings of the 6th Workshop on Geographic Information Retrieval*.

Buscaldi, D. and Rosso, P., 2008. A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science*, 22 (3), 301–313.

Chen, W., Cai, Y., Leung, H., and Li, Q., 2010. Generating ontologies with basic level concepts from folksonomies. *Procedia Computer Science*, 1 (1), 573–581.

Chiang, D., 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33 (2), 201–228.

Chowdhury, G., 2010. *Introduction to modern information retrieval*. New York: Facet publishing.

Clough, P., 2005. Extracting metadata for spatially-aware information retrieval on the internet. *In*: *Proceedings of the 2005 workshop on Geographic information retrieval*. 25–30.

Coates, R., 2006. Properhood. *Language*, 82 (2), 356–382.

Cohen, M., 1999. *The Sentimental Education of the Novel*. New Jersey: Princeton University Press.

Cooper, D. and Gregory, I.N., 2011. Mapping the English Lake District: a literary GIS. *Transactions of the Institute of British Geographers*, 36 (1), 89–108.

Couclelis, H., 2010. Ontologies of geographic information. *International Journal of Geographical Information Science*, 24 (12), 1785–1809.

Crandall, D.J., Backstrom, L., Huttenlocher, D., and Kleinberg, J., 2009. Mapping the world's photos. *Proceedings of the 18th international conference on World wide web WWW 09*, 7 (1), 761.

Dehn, M., Ga, H., and Dikau, R., 2001. Principles of semantic modeling of landform structures. *Computers & Geosciences*, 27, 1005–1010.

Deng, Y., 2007. New trends in digital terrain analysis: landform definition, representation, and classification. *Progress in physical geography*, 31 (4), 405–419.

Derungs, C., Palacio, D., and Purves, R.S., 2012. Resolving fine granularity toponyms: Evaluation of a disambiguation approach. *In*: *GIScience 2012, 7th International Conference on Geographic Information Science*.

Derungs, C. and Purves, R., 2013. From text to landscape: Locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus. *International Journal of Geographical Information Science*.

Derungs, C. and Purves, R.S., 2007. Empirical experiments on the nature of Swiss mountains. *In*: *GISRUK 2007 Geographical Information Science Research Conference*. Maynooth.

Derungs, C. and Purves, R.S., 2012. Measuring topographic similarity of toponyms. *In*: *Proceedings of the 15thAGILE International Conference on Geographic Information Science*. Avignon.

Derungs, C., Purves, R.S., and Waldvogel, B., 2011. Toponym disambiguation of landscape features using geomorphometric characteristics. *In*: *Proceedings of the 11th International Conference on GeoComputation, London, UK*. 106–110.

Derungs, C., Wartmann, F.M., Purves, R.S., and Mark, D.M., 2013. The Meanings of Generic Parts of Toponyms: Use and Limitations of Gazetteers in Studies of Landscape Terms. *Lecture Notes in Computer Science*.

Duce, S. and Janowicz, K., 2010. Microtheories for spatial data infrastructures-accounting for diversity of local conceptualizations at a global level. *In*: *Geographic Information Science*. Springer, 27–41.

Edwardes, A. and Purves, R., 2007. A theoretical grounding for semantic descriptions of place. *Web and Wireless Geographical Information Systems*, 106–120.

Edwardes, A.J., Purves, R.S., Bircher, S., and Matyas, C., 2007. *TRIPOD. TRI-Partite multimedia Object Description*. Zurich.

Van Eetvelde, V. and Antrop, M., 2009. A stepwise multi-scaled landscape typology and characterisation for trans-regional integration, applied on the federal state of Belgium. *Landscape and Urban Planning*, 91 (3), 160–170.

Egenhofer, M. and Mark, D.M., 1995. Naive Geography. *In*: *Spatial Information Theory: A Theoretical Basis for GIS*. 1–15.

Everett, C., 2013. Evidence for Direct Geographic Influences on Linguistic Sounds: The Case of Ejectives. *PloS one*, 8 (6).

Faber, V., 1994. Clustering and the continuous k-means algorithm. *Los Alamos Science*, 22, 138–144.

Fairclough, G., 2006. A new landscape for cultural heritage management: characterisation as a management tool. *Landscapes Under Pressure*, 55–74.

Fellbaum, C., 1998. A semantic network of english: the mother of all WordNets. *Computers and the Humanities*, 32 (2-3), 209–220.

Fisher, P., 2000. Sorites paradox and vague geographies. *Fuzzy Sets and Systems*, 113 (1), 7–18.

Fisher, P., Wood, J., and Cheng, T., 2004. Where is Helvellyn? Fuzziness of multi-scale landscape morphometry. *Transactions of the Institute of British Geographers*, 29 (1), 106–128.

Fisher, P.F., 1991. Modelling soil map-unit inclusions by Monte Carlo simulation. *International Journal of Geographical Information System*, 5 (2), 193–208.

Freeman, T.G., 1991. Calculating catchment area with divergent flow based on a regular grid. *Computers & Geosciences*, 17 (3), 413–422.

Frege, G., 1994. Über sinn und bedeutung. *Wittgenstein Studien*, 1 (1).

Fu, G., Jones, C.B., and Abdelmoty, A.I., 2005. Ontology-based spatial query expansion in information retrieval. *In*: *On the move to meaningful internet systems 2005: CoopIS, DOA, and ODBASE*. Springer, 1466–1482.

Gan, Q., Attenberg, J., Markowetz, A., and Suel, T., 2008. Analysis of geographic queries in a search engine log. *In*: *Proceedings of the first international workshop on Location and the web*. 49–56.

Garbin, E. and Mani, I., 2005. Disambiguating toponyms in news. *In*: *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. 363–370.

Gibson, J.J., 1979. *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin Company.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69 (4), 211–221.

Goodrum, A.A., 2000. Image Information Retrieval: An Overview of Current Research. *Informing Science*, 3 (2), 63–66.

Granö, J.G., 1997. *Pure Geography*. Baltimore: John Hopkins University Press.

Grohmann, C.H., Smith, M.J., and Riccomini, C., 2011. Multiscale analysis of topographic surface roughness in the Midland Valley, Scotland. *Geoscience and Remote Sensing, IEEE Transactions on*, 49 (4), 1200–1213.

Gruber, T., 2007a. Ontology of folksonomy: A mash-up of apples and oranges. *International Journal on Semantic Web and Information Systems*, 3 (1), 1–11.

Gruber, T., 2007b. Ontology of Folksonomy : A Mash-up of Apples and Oranges. *International Journal on Semantic Web and Information Systems*, 3 (2).

Gruber, T.R., 1993. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5 (2), 199–220.

Gschwend, C. and Purves, R.S., 2012. Exploring Geomorphometry through User Generated Content: Comparing an Unsupervised Geomorphometric Classification with Terms Attached to Georeferenced Images in Great Britain. *Transactions in GIS*, 16 (4), 499–522.

Guarino, N., 1998. Formal ontology in information systems. *In*: *FOIS'98*. Trento.

Guttman, A., 1984. *R-trees: A dynamic index structure for spatial searching*. ACM.

Haeberli, W., 2009. Gletscherschwund - Verlust eines Mythos? *Mitteilungen der Naturforschenden Gesellschaft in Bern.*, 66, 221–228.

Hard, G., 1970. Der "'Totalcharakter der Landschaft'". Re-Interpretation einiger Textstellen bei Alexander von Humboldt. *Eigene und neue Wertungen der Reisen, Arbeit und Gedankenwelt.*, 23, 49–73.

Herring, P.C., 2009. Framing Perceptions of the Historic Landscape: Historic Landscape Characterisation (HLC) and Historic Land-Use Assessment (HLA). *Scottish Geographical Journal*, 125 (1), 61–77.

Heyes, S.A., 2011. Between the trees and the tides: Inuit ways of discriminating space in a coastal and boreal landscape. *In*: D.M. Mark, A.G. Turk, N. Burenhult, and D. Stea, eds. *Landscape in Language*. New York: Berghahn Books, 187–223.

Hill, L.L., 2009. *Georeferencing: The geographic associations of information*. MIT Press.

Hochberg, J., 1978. Art and perception. *Handbook of perception*, 10, 225–258.

Höfler, S. and Piotrowski, M., 2011. Building Corpora for the Philological Study of {Swiss} Legal Texts. *Journal for Language Technology and Computational Linguistics*, 26 (2), 77–88.

Hollenstein, L. and Purves, R., 2010. Exploring place through user-generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science*, 1 (1), 21–48.

Hollink, L., Schreiber, A.T., Wielinga, B.J., and Worring, M., 2004. Classification of user image descriptions. *International Journal of Human-Computer Studies*, 61 (5), 601–626.

Hollis, J. and Valentine, T., 2001. Proper-name processing: Are proper names pure referencing expressions? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27 (1), 99.

Hotho, A., Jäschke, R., Schmitz, C., and Stumme, G., 2006. Information retrieval in folksonomies: Search and ranking. *The semantic web: research and applications*, 411–426.

Iwahashi, J. and Pike, R., 2007. Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. *Geomorphology*, 86 (3-4), 409–440.

Jackson, J.B., 1984. *Discovering the Vernacular Landscape*. New York: Yale University Press.

Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., and Stumme, G., 2007. Tag recommendations in folksonomies. *Knowledge Discovery in Databases: PKDD 2007*, 506–514.

Jett, S., 2011. Landscape embedded in language. *In*: D. Mark, A.G. Turk, and N. Burenthult, eds. *Landscape in Language*. 327–342.

Johnson, L.M. and Hunn, E.S., 2010. *Landscape ethnoecology: concepts of biotic and physical space*. Berghahn Books.

Jones, C.B. and Purves, R.S., 2008. Geographical information retrieval. *International Journal of Geographical Information Science*, 22 (3), 219–228.

Kienzle, S., 2004. The effect of DEM raster resolution on first order, second order and compound terrain derivatives. *Transactions in GIS*, 8 (1), 83–111.

Kluge, F., 2002. Etymologisches Wörterbuch der deutschen Sprache.

Kornai, A., 2006. Evaluating Geographic Information Retrieval. *Accessing Multilingual Information Repositories*, 928–938.

Kuhn, W., 2001. Ontologies in support of activities in geographical space. *International Journal of Geographical Information Science*, 15 (7), 613–631.

Kuhn, W., 2011. Ontology of landscape in language. *In*: D.M. Mark, N. Burenhult, and A.G. Turk, eds. *Landscape in Language*. 369–380.

Kupietz, M. and Keibel, H., 2009. The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. *Working papers in corpus-based linguistics and language education*, 3.

Laine-Hernandez, M. and Westman, S., 2006. Image semantics in the description and categorization of journalistic photographs. *Proceedings of the American Society for Information Science and Technology*, 43 (1), 1–25.

Lakoff, G. and Johnson, M., 1980. *Metaphors we live by*. Chicago London.

Larson, 2011. Ranking Approaches for GIR. *In*: R. Purves and C. Jones, eds. *Letters on Geographic Information Retrieval*. ACM Sigspatial Special, 37–42.

Larson, R.R. and Frontiera, P., 2004. Spatial ranking methods for geographic information retrieval (GIR) in digital libraries. *Research and Advanced Technology for Digital Libraries*, 45–56.

Leidner, J.L., 2004. Toponym Resolution in Text : " Which Sheffield is it ?" *Proceedings of the 27th annual international ACM conference on Research and development in information retrieval*.

Leidner, J.L., 2007. Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names. *Evaluation*. University of Edinburgh.

Leidner, J.L. and Lieberman, M.D., 2011. Detecting geographical references in the form of place names and associated spatial natural language A Processing Model For Textually Encoded Geo-. *Machine Learning*, 1–7.

Leveling, J. and Veiel, D., 2007. Experiments on the exclusion of metonymic location names from GIR. *In*: *Evaluation of Multilingual and Multi-modal Information Retrieval*. Springer, 901–904.

Levinson, S.C., 2011. Foreword. *In*: D.M. Mark, A.G. Turk, N. Burenthult, and D. Stea, eds. *Landscape in Language*. New York: Berghahn Books, ix–x.

Li, H., Srihari, R.K., Niu, C., and Li, W., 2003. InfoXtract location normalization: a hybrid approach to geographic references in information extraction. *In*: *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references-Volume 1*. 39–44.

Li, Z., Wang, C., Xie, X., Wang, X., and Ma, W.-Y., 2006. Indexing implicit locations for geographical information retrieval. *GIR. Department of Geography, University of Zurich*.

Lieberman, M.D., Samet, H., Sankaranarayanan, J., and Sperling, J., 2007. STEWARD: architecture of a spatio-textual search engine. *In*: *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*.

Lock, G., 2010. Representations of space and place in the humanities. *The Spatial Humanities. GIS and the Future of Humanities Scholarship*, 89–108.

Mandelbrot, B.B., 1967. How long is the coast of Britain. *Science*, 156 (3775), 636–638.

Mandl, 2011. Evaluating GIR: Geography-oriented or User-oriented? *In*: R. Purves and C. Jones, eds. *Letters on Geographic Information Retrieval*. ACM Sigspatial Special, 42–46.

Mandl, T., Gey, F., Di Nunzio, G., Ferro, N., Larson, R., Sanderson, M., Santos, D., Womser-Hacker, C., and Xie, X., 2008. Geoclef 2007: the clef 2007 cross-language geographic information retrieval track overview. *Advances in Multilingual and Multimodal Information Retrieval*, 745–772.

Mani, I., Hitzeman, J., and Clark, C., 2008. SpatialML: Annotation Scheme, Corpora, and Tools. *In*: *The Workshop Programme Methodologies and Resources for Processing Spatial Language*.

Manning, C.D., Raghavan, P., and Schütze, H., 2008. *Introduction to information retrieval*. Cambridge University Press Cambridge.

Marcus, M.P., Marcinkiewicz, M.A., and Santorini, B., 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19 (2), 313–330.

Mark, D.M., Turk, A., and Stea, D., 2007. Progress on Yindjibarndi ethnophysiography. *Spatial information theory*, 1–19.

Mark, D.M. and Turk, A.G., 2003. Landscape Categories in Yindjibarndi : Ontology , Environment , and Language. *Language*, (1970).

Mark, D.M., Turk, A.G., Burenthult, N., and Stea, D., 2011. *Landscape in Language*. New York: Berghahn Books.

Mark, D.M., Turk, A.G., and Stea, D., 2010. Ethnophysiography of Arid Lands. *Landscape Ethnoecology: Concepts of Biotic and Physical Space*, 27.

Marr, D., 1982. Vision: A computational investigation into the human representation and processing of visual information, Henry Holt and Co. *Inc., New York, NY*.

Martins, B., Anastácio, I., and Calado, P., 2010. A Machine Learning Approach for Resolving Place References in Text. *Machine Learning*, 221–236.

Michel, J.-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., and Aiden, E.L., 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331 (6014), 176–182.

Miller, H.J., 2010. The Data Avalanche is here. Shouldn't we be digging? *Journal of Regional Science*, 50 (1), 181–201.

Moore, I.D., Grayson, R.B., and Ladson, A.R., 2006. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. *Hydrological processes*, 5 (1), 3–30.

Moretti, F., 1998. *Atlas of the European Novel: 1800-1900*. Verso.

Moretti, F., 2007. Graphs, Maps, Trees: Abstract Models for Literary History. *New Left Review*, 68 (1), 132–135.

Mücher, C.A., Klijn, J.A., Wascher, D.M., and Schaminée, J.H.J., 2010. A new European Landscape Classification (LANMAP): A transparent, flexible and user-oriented methodology to distinguish landscapes. *Ecological Indicators*, 10 (1), 87–103.

Muir, J., 1917. *The story of my boyhood and youth*. Houghton Mifflin.

Müller, G., 1977. Zur Geschichte des Wortes Landschaft. *„Landschaft "als interdisziplinäres Forschungsproblem*, 4–12.

Murton, B., 2011. "Mirror knowledge" and "simultaneous landscapes" among Maori. *In*: D.M. Mark, A.G. Turk, N. Burenhult, and D. Stea, eds. *Landscape in Language*. New York: Berghahn Books, 73–100.

Nature, 2007. A matter of trust. *Nature*, (449), 637–638.

Naveh, Z. and Lieberman, A.S., 1984. *Landscape Ecology: Theory and Application*. Springer.

Nelson, K., Hampson, J., and Shaw, L.K., 1993. Nouns in early lexicons: evidence, explanations and implications. *Journal of Child Language*, 20 (01), 61–84.

O'Sullivan, D. and Unwin, D.J., 2003. *Geographic information analysis*. John Wiley & Sons.

Overell, S. and Rüger, S., 2008. Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science*, 22 (3), 265–287.

Van Overschelde, J., 2004. Category norms: An updated and expanded version of the norms. *Journal of Memory and Language*, 50 (3), 289–335.

Palacio, D., Cabanac, G., Sallaberry, C., and Hubert, G., 2010. Measuring Effectiveness of Geographic IR Systems in Digital Libraries. *Research and Advanced Technology for Digital Libraries*, 340–351.

Piatti, B., 2008. *Die Geographie der Literatur: Schauplätze, Handlungsräume, Raumphantasien*. Wallstein.

Pickles, J., 1994. *Ground truth: The social implications of geographic information systems*. The Guilford Press.

Pike, R.J., Evans, I.S., and Hengl, T., 2009. Geomorphometry: A Brief Guide. *Terrain*, 33.

Piotrowski, M., 2012. *Natural Language Processing for Historical Texts*. San Rafael, CA, USA: Morgan & Claypool.

Purves, R.S., Clough, P., Jones, C.B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A.K., Vaid, S., and Yang, B., 2007. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science*, 21 (7), 717–745.

Purves, R.S., Edwardes, A.J., and Wood, J., 2011. Describing place through user generated content. *First Monday*, 16 (9).

Purves, R.S. and Jones, C.B., 2011. *Letters on Geographic Information Retrieval*. SIGSpatial.

Raper, J., 2007. Geographic relevance. *Journal of Documentation*, 63 (6), 836–852.

Rattenbury, T. and Naaman, M., 2009. Methods for extracting place semantics from Flickr tags. *ACM Transactions on the Web*, 3 (1), 1–30.

Reitman, W.R., 1965. Cognition and thought: an information processing approach.

Rosch, E., 1973. Natural Categories. *Cognitive Psychology*, 4 (3), 328–350.

Rosch, E. and Lloyd, B.B., 1978. Principles of categorization. *In*: *Cognition and categorization*. Erlbaum.

Samet, H., 2006. *Foundations of multidimensional and metric data structures*. Morgan Kaufmann.

Sauerland, U., 2011. Vagueness in language: the case against fuzzy logic revisited. *Reasoning under Vagueness-Logical, Philosophical, and Linguistic Perspectives, Studies in Logic series of College Publications*.

Saur, C.O., 1913. The Morphology of Landscape. *University of California Publications in Geography*, 2 (2), 19–53.

Sennrich, R., Schneider, G., Volk, M., and Warin, M., 2009. A new hybrid dependency parser for German. *In*: *Proceedings of GSCL-Conference*. Potsdam.

Serdyukov, P., Murdock, V., and Van Zwol, R., 2009. Placing flickr photos on a map. *In*: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 484–491.

Shaftesbury, A.A.C., 1964. Earl of " The Moralists." *Characteristics of Men, Morals, Opinions and Times*.

Shatford, S., 1986. Analyzing the Subject of a Picture: A Theoretical Approach. *Cataloging Classification Quarterly*, 6 (3), 39–62.

Simmel, G., 1913. Philosophie der Landschaft. *Eine bremische Monatsschrift*, 3 (2).

Sinha, G. and Mark, D.M., 2010. Cognition-based extraction and modelling of topographic eminences. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 45 (2), 105–112.

Smith, B., 1995. Formal ontology, common sense, and cognitive science. *International Journal of Human Computer Studies*, 43, 641–667.

Smith, B., 2003. Ontology. *The Blackwell guide to the philosophy of computing and information*, 153–166.

Smith, B., 2007. On Drawing Lines on a Map, (1995), 475–484.

Smith, B. and Mark, D.M., 1998. Ontology and geographic kinds. *In*: T.K. Poiker and N. Chrisman, eds. *Proceedings of the 8th International Symposium on Spatial Data Handling (SDH'98)*. 308–320.

Smith, B. and Mark, D.M., 2001. Geographical categories: an ontological investigation. *International Journal of Geographical Information Science*, 15 (7), 591–612.

Smith, B. and Mark, D.M., 2003. Do mountains exist? Towards an ontology of landforms. *Environment and Planning B: Planning and Design*, 30 (3), 411–427.

Smith, D.A. and Crane, G., 2001. Disambiguating Geographic Names in a Historical Digital Library. *In*: *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*. 127–136.

Steyvers, M. and Griffiths, T., 2007. Probabilistic topic models. *Handbook of latent semantic analysis*, 427 (7), 424–440.

Straumann, R., 2010. Extraction and Characterisation of Landforms from Digital Elevation Models: Fiat Parsing the Elevation Field. PhD thesis. University of Zurich, Switzerland.

Straumann, R. and Korup, O., 2009. Quantifying postglacial sediment storage at the mountain-belt scale. *Geology*, 37 (12), 1079–1082.

Straumann, R. and Purves, R., 2008. Delineation of valleys and valley floors. *Geographic Information Science*, 320–336.

Stremlow, M. and Sidler, C., 2002. *Schreibzüge durch die Wildnis: Wildnisvorstellungen in Literatur und Printmedien der Schweiz*. Haupt.

Tarboton, D.G., Bras, R.L., and Rodriguez-Iturbe, I., 1991. On the extraction of channel networks from digital elevation data. *Hydrological processes*, 5 (1), 81–100.

Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. *Economic geography*, 46, 234–240.

Topaha, C., 2011. Navajo landscape and its contexts. *In*: D.M. Mark, A.G. Turk, N. Burenhult, and D. Stea, eds. *Landscape in Language*. Amsterdam: John benjamins Publishing Company, 343–353.

Tuan, Y., 1974. *Topophilia. A study of Environmental Perception, Attitudes, and Values*. New Jersey: Prentice-Hall Inc.

Turk, A., Mark, D.M., and Stea, D., 2011. Ethnophysiography. *In*: D.M. Mark, A.G. Turk, N. Burenhult, and D. Stea, eds. *Landscape in Language*. New York: Berghahn Books, 25–45.

Tversky, B. and Hemenway, K., 1983. Categories of Environmental Scenes. *Cognitive Psychologies*.

Vaid, S., Jones, C.B., Joho, H., and Sanderson, M., 2005. Spatio-textual indexing for geographical search on the web. *In*: *Advances in Spatial and Temporal Databases*. Springer, 218–235.

Vale, T.R., 2002. *Fire, native peoples, and the natural landscape*. Island Press.

Volk, M., Bubenhofer, N., Althaus, A., Bangerter, M., Furrer, L., and Ruef, B., 2010. Challenges in building a multilingual alpine heritage corpus. *In*: *Proceedings of LREC*. Malta.

Volk, M. and Steinhardt, U., 2002. The landscape concept. What is a landscape? *In*: O. Bastian and U. Steinhardt, eds. *Development and Perspectives in Landscape Ecology*. Dordrecht: Kluwer Academic Publishers.

Voorhees, E., Harman, D.K., and others, 2005. *TREC: Experiment and evaluation in information retrieval*. MIT press Cambridge.

Vander Wal, T., 2007. Folksonomy [online]. Available from: http://vanderwal.net/folksonomy.html.

Walter, F., 1996. *Bedrohliche und bedrohte Natur – Umweltgeschichte der Schweiz seit 1800*. Zürich: Chronos Verlag.

Warhig, R., 1994. *Deutsches Wörterbuch*. Bertelsmann Lexikon.

White, R. and Buscher, G., 2012. Characterizing local interests and local knowledge. *In*: *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*. 1607–1610.

Williamson, T., 1996. *Vagueness*. Routledge.

Wing, B. and Baldridge, J., 2011. Simple supervised document geolocation with geodesic grids. *In*: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 955–964.

Wittgenstein, L., 1922. *Tractatus logico-philosophicus*. London: Kegan Paul.

Wood, J., 1996. The geomorphological characterisation of digital elevation models. University of Leicester.

Wood, J., Dykes, J., Slingsby, A., and Clarke, K., 2007. Interactive visual exploration of a large spatio-temporal dataset: reflections on a geovisualization mashup. *IEEE Transactions on Visualization and Computer Graphics*, 13 (6), 1176–1183.

Woodruff, A.G. and Plaunt, C., 1994. GIPSY: Automated Geographic Indexing of Text Documents. *Journal of the American Society for Information Science*, 45 (9), 645–655.

Worster, D., 2008. Environmentalism Goes Global. *Diplomatic History*, 32 (4), 639–641.

Wu, H.C., Luk, R.W.P., Wong, K.F., and Kwok, K.L., 2008. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, 26 (3), 13.

Wylie, J., 2009. Landscape, absence and the geographies of love. *Transactions of the Institute of British Geographers*, 34 (3), 275–289.

Younis, E.M.G., Jones, C.B., Tanasescu, V., and Abdelmoty, A.I., 2012. Hybrid Geo-spatial Query Methods on the Semantic Web with a Spatially-Enhanced Index of DBpedia. *In*: *GIScience*. 340–353.

Zadeh, A.L., 1965. Fuzzy Sets. *Information and Control*, 8, 338–353.

Zedler, J.H., 1749. *Universal-Lexikon*. Leipzig: Zedler.

Zipf, G.K., 1935. *The psycho-biology of language: an introduction to dynamic philology*. Boston: Houghton Mifflin company.

Index of Figures

Index of Tables

Appendix

Appendix A

Appendix A. Annotation rules for identifying natural features from lists of nouns.

German Version

Das Ziel dieser Aufgabe ist das Annotieren von Nomen als natürliche Objekte. Natürliche Objekte müssen dabei von allen anderen Arten von Nomen unterschieden werden. Die folgende Liste enthält einige Regeln die das Annotieren von besonders schwierigen Fällen erleichtern soll. Oft ist der erste Eindruck aber aussagekräftig.

Die Annotation wird in der Spalte ‚nat. Objekt' gemacht (in der Tabelle top1500Nouns_textBerg.xlsx). Es wird nur zwischen natürlichem Objekt („1") und allen anderen Nomen („", nichts) unterschieden, Fragezeichen und Kommentare sind keine gültigen Annotationen. Jedes Vorkommen eines Nomens muss annotiert werden, unabhängig davon ob an früherer Stelle das gleiche Nomen bereits in unterschiedlicher Deklination vorgekommen ist (z.B. Berg, Berge, Bergen).

Annotations Regeln

Natürliche Objekte sind…

…*generisch*. Das heisst, dass sie eine Objekt-Klasse vertreten und nicht individuelle Objekte. Berg (ok) ist ein natürliches Objekt, Matterhorn (nicht ok) and Alpen (nicht ok) nicht, sie sind Individuen.

…*natürlich*. ‚Natürlichkeit' ist manchmal eine schwierige und nicht eindeutig feststellbare Eigenschaft. Für diese Annotation bedeutet ‚natürlich', dass die Materie des Objektes nicht massgeblich vom Menschen geschaffen oder transformiert wurde. Alp (ok) ist natürlich (obwohl kultiviert handelt es sich noch immer um Wiesen), Alphütten (nicht ok) sind künstlich (Wände, Dach und Boden bestehen aus Materialien die transportiert und stark bearbeitet werden mussten um eine Hütte damit zu bauen). Man kann sich auch die Frage stellen ob eine menschliche Aktivität nötig ist deren einziger Zweck die Erstellung dieses Objektes ist. Falls ja handelt es sich um ein künstliches Objekt (bauen einer Alphütte (nicht ok) ist eine Aktivität die eigens der Erschaffung einer Hütte dient, wandern ist eine Aktivität die nicht das primäre Ziel hat einen Pfad (ok) zu erschaffen).

…*keine Aktivitäten*. Manchmal können Nomen Aktivitäten und natürliche Objekte sein. In diesen Fällen entscheiden wir uns für Aktivität und das entsprechende Nomen wird nicht Annotiert. Aufstieg (nicht ok) ist eine Aktivität die ebenfalls ein ‚natürliches' Objekt bezeichnen kann. Eine Entscheidungshilfe ist, falls ein Nomen in direkter Beziehung zu einem Verb, mit der gleichen Bedeutung, steht (Aufstieg -> aufsteigen) wird es nicht als natürliches Objekt annotiert. Das gilt nicht falls sich bei der Umformung in ein Verb der Sinn ändert (Berg (ok) -> bergen).

…kein Phänomen oder Qualität. Natürliche Objekte sind weitgehend unabhängige Existenzen. Im Gegensatz dazu sind Phänomene oft nur Spezifikationen von natürlichen Objekten. Schnee (nicht ok) oder Eis (nicht ok) werden oft verwendet um den Zustand von Bergen (ok) näher zu beschreiben. Ein Schneefeld (ok) hingegen ist ein unabhängiges natürliches Objekt.

Appendix B

Appendix B. List of all natural features identified from the 1500 most frequent nouns in the Text+Berg corpus. Applied are the counts of these natural features as resolved from the whole corpus (*count T+B*).

rank	nat. features	count T+B	rank	nat. features	count T+B	rank	nat. features	count T+B
1	gipfel	29635	36	baum	1955	71	felsgrat	882
2	berg	27037	37	flanke	1784	72	gipfelgrat	861
3	alp	24840	38	südwand	1768	73	schutthalde	833
4	gletscher	17849	39	weide	1710	74	westwand	810
5	fels	17522	40	schneefeld	1687	75	steilhang	792
6	grat	14337	41	fluss	1653	76	paß	787
7	wand	14202	42	geröll	1645	77	vorgipfel	754
8	tal	10273	43	ostgrat	1608	78	kuppe	753
9	spitze	6544	44	horn	1590	79	gletscherzunge	747
10	thal	5705	45	wiese	1567	80	südostgrat	727
11	stein	5626	46	westgrat	1514	81	talboden	722
12	hang	5551	47	nordgrat	1511	82	nordostgrat	691
13	wald	5199	48	abgrund	1429	83	nordflanke	670
14	see	4967	49	felsblock	1405	84	südwestgrat	666
15	gebirge	4822	50	abhang	1386	85	küste	630
16	platte	4078	51	südgrat	1386	86	alpweide	593
17	gestein	3717	52	überhang	1385	87	wüste	558
18	landschaft	3614	53	bergschrund	1364	88	einzugsgebiet	551
19	pass	3580	54	loch	1364	89	nordwestgrat	527
20	schlucht	3418	55	schrund	1335	90	westflanke	526
21	spalte	3345	56	plateau	1319	91	waldgrenze	515
22	felswand	3169	57	massiv	1308	92	südflanke	511
23	bach	3103	58	insel	1269	93	talseite	487
24	scharte	2800	59	wasserfall	1187	94	wasserscheide	486
25	gelände	2662	60	passhöhe	1167	95	ostflanke	474
26	meer	2637	61	hauptgipfel	1118			
27	pfad	2610	62	feld	1097			
28	kamm	2585	63	schutt	1069			
29	hochgebirge	2479	64	ostwand	1060			
30	rinne	2477	65	matten	1060			
31	moräne	2312	66	eiswand	1044			
32	nordwand	2217	67	blume	950			
33	ebene	2074	68	gebirgswelt	911			
34	sattel	2049	69	hügel	909			
35	quelle	2009	70	terrain	894			